Metagenomic approaches to microbial source tracking

A Thesis

submitted in partial fulfilment

of the requirements for the Degree

of

Master of Science in Cellular and Molecular Biology

at the

University of Canterbury

by Carina L. Davis

University of Canterbury

2013

Dedicated to my little Zachary Quack,
who after five years on this journey with me
is not so little anymore.

I hope that one day you will enjoy science
as much as I do.

Table of Contents

Acknowledgements

Without the assistance of a vast number of people, I would not have been able to undertake the journey which has led to the completion of this thesis, and I would like to thank the following people for their huge contribution.

First and foremost, I would like to thank both of my supervisors, Dr Brent Gilpin (ESR) and Dr Arvind Varsani (University of Canterbury). Without your guidance, support, and encouragement, I would have been lost before I even started. Thank you for all your patience as I continued to try to come to grips with this ever-changing field, and for all your feedback throughout the whole process.

I would like to acknowledge ESR for funding this project, and thank all the staff at CSC for making me so welcome over the past couple of years. In particular, I would like to thank the Water Molecular Biology group for all your help and support. To Beth Robson and Paula Scholes, thank you for all your technical assistance and for continuing to answer all my questions in the lab.

A huge thanks must go to the virus lab group at Canterbury University. For the friendship and laughs, and for making me feel part of the group, even when I didn't really know what you were talking about most of the time. It's been fantastic being a part of such an awesome group of people.

To all the inhabitants of the ESR portacom, regardless of how long your stay has been, you have all been involved in this journey, and I thank you for your support and entertainment. You have all been a delight to share an office with, and I wish you all well on your future scientific ventures. Special thanks must go to Theresa Stotesbury, who kept me sane in the first few months, and to Niki Osborne, for all her motivation and advice during the last few.

There are so many friends and family who have supported me over the past five years. Without you, I would never have been able to get this far, and I am forever grateful. Special thanks must go to both sets of parents, Mum and Dad D, and Mum and Dad S, as well as Hayley and Kelvin, and George and Debbie. Thank you for the countless hours of babysitting, glasses of wine, and continued support.

And finally, to my family, Corey and Zac, who have kept me going through it all. From all the morning coffees, to providing the solutions when all I wanted was to throw the computer out the window, and everything in between, I couldn't have managed without the support of my partner Corey. Thank you for putting up with my long hours over the past few months in particular, and ensuring that everything else continued to run smoothly. Zac, you continue to be a charmer, and have kept me smiling throughout. Thank you for all your hugs and kisses, and for letting Mummy away with not playing games so she could finish writing her book.

List of Figures

Chapter One

Chapter Two

Chapter Three

List of Tables

Chapter One

Chapter Two

Chapter Three

Chapter Four

List of Abbreviations

°C	degrees Celsius
%	percent
A	adenine
A_{260}	absorbance at 260 nanometres
A_{280}	absorbance at 280 nanometres
AFLP	amplified fragment length polymorphism
ARA	antibiotic resistance analysis
ATP	adenosine triphosphate
biom	biological observation matrix
bp	base pair(s)
C	cytosine
C1 – 10	conserved region of the 16S rRNA gene
CCD	charge-coupled device
cDNA	complementary deoxyribonucleic acid
Cp	threshold cycle (crossing point)
CUP	carbon utilisation profiling
ddNTP	dideoxy nucleotide triphosphate
DGGE	denaturing gradient gel electrophoresis
dH_2O	distilled water
DNA	deoxyribonucleic acid
dNTP	deoxy nucleotide triphosphate
ds	double stranded
EDTA	ethylenediamine tetra-acetic acid
emPCR	emulsion polymerase chain reaction

ESR	Institute of Environmental Science and Research Limited
EtBr	ethidium bromide
FAME	fatty acid methyl esters
FST	faecal source tracking
G	guanine
g	gram(s)
GB	gigabyte
Gb	giga-base pairs
GITC	guanidine isothiocyanate
GS	Genome Sequencer
h	hour(s)
HiFi	high fidelity
ID	identification
Inc.	Incorporated
Kb	kilo-base pairs
LDM	library-dependent method
LH-PCR	length heterogeneity polymerase chain reaction
LIM	library-independent method
log	logarithm to the base 10
M	molar
Mb	mega-base pairs
MEGAN	MetaGenome Analyzer
MID	multiplex identifier
min	minute(s)
ml	millilitre(s)
mm	millimetre(s)
mM	millimolar

mRNA	messenger ribosenucleic acid
MST	microbial source tracking
mtDNA	mitochondrial deoxyribonucleic acid
MUSCLE	Multiple Sequence Comparisons by Log-Expectation
ng	nanogram(s)
NGS	next generation sequencing
no.	number
nt	nucleotide
NZGL	New Zealand Genomics Limited
OTU	operational taxonomic unit
PC	principal coordinate
PCoA	principal coordinate analysis
PCR	polymerase chain reaction
PFGE	pulse-field gel electrophoresis
PGM	Personal Genome Machine
pmol	picomole(s)
PPi	pyrophosphate
PTP	PicoTiterPlate
PyNAST	Python Nearest Alignment Space Termination
QIIME	Quantitative Insights Into Microbial Ecology
qPCR	quantitative polymerase chain reaction
RAM	random access memory
RDP	Ribosomal Database Project
Rep-PCR	repetitive element sequencing-based polymerase chain reaction
RNA	ribonucleic acid
rpm	revolutions per minute
rRNA	ribosomal ribonucleic acid

s	second(s)
S.D.	standard deviation
sff	standard flowgram format
SMRT	single molecule real time
SOLiD	Support Oligonucleotide Ligation Detection
spp.	species
ss	single stranded
T	thymine
Taq	*Thermus aquaticus*
TBE	tris-borate EDTA buffer
TCEP	tris (2-carboxyethyl) phosphine
TE	tris EDTA buffer
TGS	third generation sequencing
T_m	melting point
TMDL	total maximum daily load
t-RFLP	terminal restriction fragment length polymorphism
µl	microliter(s)
µM	micromolar
UniFrac	unique fraction metric
UPGMA	unweighted pair group method with arithmetic mean
US EPA	United States Environmental Protection Agency
UV	ultraviolet
V	volts
V1 – 9	hypervariable region of the 16S rRNA gene
vs.	verses
ZMW	zero-mode waveguide

Abstract

Water sources are susceptible to faecal contamination from animal and human pollution sources. Pollution of our waterways has significant implications on human health, especially from a pathogen perspective. Microbial source tracking (MST) is a promising field which aims to identify the sources of faecal contamination, and thereby allowing for the development of effective management strategies to minimise pollution and the impact on human health. Many of the currently used methods rely on the identification of host-specific markers within the 16S ribosomal RNA (rRNA) gene of bacteria by use of amplification techniques such as polymerase chain reaction (PCR). However, these methods can be limited by sensitivity, quantification, geographical differences and issues of cost which can limit how many markers are evaluated.

Developments in DNA sequencing technologies over the past decade have led to a number of next generation sequencing (NGS) platforms which have a rapid, high throughput approach, resulting in an exponential decrease in the cost of sequencing. This has enabled the development of sequence-based metagenomics, where entire communities from environmental samples can be analysed based on their genetic material. The ability to barcode allows for analysis of multiple samples at once, reducing the cost of sequencing environmental samples even further. This is a promising technique for MST, which has had little investigation to date.

The primary focus of the studies described in this thesis was to evaluate the use of NGS technology through a metagenomic approach. Roche 454 amplicon sequencing was used to sequence a 16S rRNA gene target, amplified from faecal and water samples from various sources in New Zealand. Barcode strategies were incorporated in the amplification design to allow multiple samples to be sequenced simultaneously. A proof-of-concept study initially utilised a small sequence dataset to evaluate a range of analysis tools available. Taxonomic identification and diversity measures were used to evaluate a selection of currently available tools designed for analysing metagenomic data, with the Quantitative Insights Into Microbial Ecology (QIIME) platform decided upon for further studies. A larger study, including 35 faecal samples from 13 difference sources and 10 water samples, resulted in 522,065 raw sequencing reads. Diversity results suggest three phyla, *Bacteroidetes, Firmicutes* and *Proteobacteria,* are strongly represented across all faecal sources analysed. Microbial diversity analysis using clustering techniques provided evidence of host source being the largest influence on bacterial diversity, with samples from each source generally clustering together. This technique could not be used to identify sources of contamination sources in water samples as the water samples all clustered separately from the faecal samples. More successful was the use of taxonomic classifications to determine bacteria genera that were potentially specific to one source. Water samples were screened for these genera, with six out of the ten water samples being indicators of either ruminant or human

contamination. Faecal and water samples were also analysed for a selection of published 16S rRNA PCR markers, using a computational motif-based search method. Of the twenty motifs screened for, 14 were found to be relatively source-specific for ruminant, human, dog or pig faecal samples, with some cross-reactivity with chicken and possum samples. Using this method, the contamination source for six of the ten water samples was identified, with the remaining four samples found to not have enough sequences to assess with confidence. Both metagenomic strategies produced comparable results which were consistent with previous MST analysis.

This project demonstrates the potential application of next generation sequencing technologies to microbial source tracking, suggesting the possibility this approach to replace existing microbial source tracking methods.

Chapter One
Introduction

1.1 Microbial source tracking

Faecal microorganisms are one of the primary pollutants of water throughout the world (Santo Domingo *et al.*, 2007). Although the frequency and level of severity are higher in developing countries, waterborne outbreaks are common in all countries, with global estimates in 2003 suggesting an excess of 175 million cases of infectious diseases each year (Shuval, 2003). The majority of bacterial, viral and protozoan pathogens associated with waterborne outbreaks are primarily found in the faeces of higher mammals (Leclerc *et al.*, 2002), therefore, preventing mammalian faecal contamination of waterways is of critical importance for human health, as exposure to faecal polluted water can cause gastroenteritis, as well as respiratory and eye-, ear- and skin-related illnesses (Prüss, 1998; Stewart *et al.*, 2007). Avian species are also important sources of pathogens, such as *Cryptosporidium parvum, Giardia spp.* and *Campylobacter spp.* (Graczyk *et al.*, 1998; Kassa *et al.*, 2004; Moriarty *et al.*, 2011). As the general public have become increasingly aware of the potential health risks associated with faecal contaminated water, there has been an increased frequency of water quality monitoring across the world.

The term ‘microbial source tracking’ (MST), describes a variety of phenotypic and genotypic methods used to determine sources of faecal contamination in water. The term ‘faecal source tracking’ (FST) has also been used, as this does not specifically imply the use of microbes as the detection method (Field and Samadpour, 2007). FST includes chemical methods of source detection, such as faecal sterols and fluorescent whitening agents (Sinton *et al.*, 1998). MST is a field in its infancy with the primary goal to develop tools that determine the host origins of enteric microorganisms found in waterways and other environmental samples (Stewart *et al.*, 2007), and has evolved rapidly since the first approaches were introduced in 1995 (Hagedorn and Liang, 2011). The general hypothesis of MST is that some microorganisms have an exclusive or preferential association with the gastrointestinal track of a particular host species, and that these host-specific microorganisms are shed in faeces, which can then be detected in water bodies (Harwood and Stoeckel, 2011).

Faecal indicator bacteria have been used for over a century to identify faecal contamination in water (Harwood, 2007; Leclerc *et al.*, 2001), but development of alternative methods has been required in order to detect the origin of faecal pollution. There are now a number of methods in use for the identification of specific characteristics associated with faeces which can be used to identify the host source. These can generally be divided into two groups: library-dependent methods (LDMs) and library-independent methods (LIMs). LDMs rely on the construction of a large library of bacterial profiles from known faecal sources, isolated from a geographically defined environment. The known isolates are compared with those from the contaminated matrix of interest to determine the source of pollution. In comparison, LIMs rely on the detection of a particular host-specific organism or gene in the contaminated matrix, typically using molecular methods such as polymerase chain reaction (PCR).

Each method has its own advantages and disadvantages, and recent recommendations have suggested a "toolbox" approach, where multiple methods are included in each study (McLellan, 2004; Plummer and Long, 2009; Stewart *et al.*, 2003). Approaches used in MST are also relevant to other research fields, such as food safety, agriculture and veterinary microbiology (Santo Domingo *et al.*, 2007).

The detection of the origin of faecal pollution in our waterways is beginning to take a prominent place in hazard identification and risk management policies. Without knowledge of contamination sources, it is difficult to conduct risk assessments, choose effective remediation strategies, or return chronically polluted waters to an acceptable quality. Knowing the source of the faecal contamination is important for assessing public health risks and for deciding what management strategies are required (Teaf and Garber, 2011). The United States Environmental Protection Agency (US EPA) requires states to establish total maximum daily loads (TMDL) for waters deemed impaired, to establish the maximum pollutant load that a water body can receive and still meet water quality standards (Santo Domingo *et al.*, 2007; Simpson *et al.*, 2002). Monitoring by MST methods provides the data needed to determine the quality of our waters, and to identify the pollutant sources. However, this is a challenging task, and methods are continuing to be developed to apply MST for monitoring, assessment and hypothesis testing (Meays *et al.*, 2004).

1.1.1 Faecal indicator bacteria

Faecal indicator bacteria, such as *Escherichia coli* (*E. coli*), *enterococci spp.* and culturable coliforms have been routinely used in monitoring and regulation of microbial contamination in waterways. They have proven to be a practical and efficient measure of contamination, contributing to vast improvements in water safety management (Hagedorn *et al.*, 2011a; Tallon *et al.*, 2005). Indicator bacteria are usually not pathogens themselves, but are bacteria assumed to be associated with faecal contamination (Meays *et al.*, 2004; Stewart *et al.*, 2007). Indicators circumvent the need to assay for every pathogen that may be present in water, and are easier and less costly to detect and quantify than pathogens (Meays *et al.*, 2004). However, indicator bacteria do not identify the source of the contamination since they are found in the faeces of a variety of warm- and cold-blooded animals (Field and Samadpour, 2007). They have also been found in several environmental sources, including soils and sediments, algal wrack and beach sands (Boehm, 2007; Boehm *et al.*, 2009; Yamahara *et al.*, 2007).

Faecal indicator bacteria have other potential limitations associated with their application, mainly related to how well they correlate with pathogens. Differences in the survival rates in water bodies, ability to multiply in water, and susceptibility to disinfection processes, can mean that these indicators may not correlate particularly well with pathogens in faecal contamination that has travelled long distances or is aged (Savichtcheva *et al.*, 2007).

1.1.2 Library-dependent methods

LDMs require the construction of a library, or database, of known faecal source profiles and characteristics that are used for comparisons against environmental isolates to determine the source of the contamination. LDMs can make use of phenotypic and genotypic characteristics. The methods rely on isolate-by-isolate typing of bacteria cultured from various faecal sources and water samples, with the isolates correlated by direct subtype matching (Harwood *et al.*, 2003; Meays *et al.*, 2006), or by statistic means (Dombek *et al.*, 2000; Hagedorn *et al.*, 1999; Harwood *et al.*, 2000; Harwood *et al.*, 2003; Ritter *et al.*, 2003). This approach assumes that a relative proportion of the characteristics in each faecal source remain constant in the environment over time. The majority of the methods utilise characteristics of faecal indicator bacteria, including *E. coli* (Anderson *et al.*, 2006; Buchan *et al.*, 2001; Casarez *et al.*, 2007a; Casarez *et al.*,

2007b; Dombek *et al.*, 2000; Guan *et al.*, 2002; Lu *et al.*, 2004; McLellan *et al.*, 2001; Vogel *et al.*, 2007); *enterococci* (Booth *et al.*, 2003; Choi *et al.*, 2003; Genthner *et al.*, 2005; Hagedorn *et al.*, 2003); faecal coliforms (Duran *et al.*, 2006; Harwood *et al.*, 2000; Haznedaroğlu *et al.*, 2005) and faecal *streptococci* (Hagedorn *et al.*, 1999; Harwood *et al.*, 2000). However, some studies have suggested that the variation of enteric bacteria is too great to be able to use as a suitable source maker (Gordon, 2001; Gordon *et al.*, 2002; Simpson *et al.*, 2002).

Effective source tracking using known-source libraries requires that the library adequately represent diversity of source faeces in the study area (Stoeckel *et al.*, 2004). Therefore, it requires generation and validation of library profiles for each new geographical location, which can be laborious and costly. For example, the genetic diversity of *E. coli* has been demonstrated to result in the need for an extremely large library, with thousands of isolates to include the majority of the potential profiles of the organism (Mott and Smith, 2011). However, limited information is available to guide the number of isolates required for a library that represents source populations of different sizes (Jenkins *et al.*, 2003; Stoeckel *et al.*, 2004; Wiggins *et al.*, 2003), especially as the size of the library needed is also dependent on the method and organism being studied (Mott and Smith, 2011; Wiggins *et al.*, 2003). Statistical analysis is often required for interpreting and identifying the sources of faecal contamination, with multiple algorithms currently used, including discriminant analysis, average similarity and k-means nearest neighbour. No single statistical approach has been found to be superior (Mott and Smith, 2011; Ritter *et al.*, 2003). The validity of results generated using library-dependent methods has also been questioned, following a number of large scale blind-sample proficiency validations (Griffith *et al.*, 2003; Harwood *et al.*, 2003; Myoda *et al.*, 2003; Stoeckel *et al.*, 2004).

1.1.2.1 Phenotypic methods

Phenotypic methods rely on the products of gene expression, comparing patterns produced when bacteria isolates are subjected to a range of antibiotics, or grown on different media sources (Griffith *et al.*, 2003; Tallon *et al.*, 2005). Methods include antibiotic resistance analysis (ARA), carbon utilisation profiling (CUP), and fatty acid methyl esters (FAME). Of these methods, ARA is the most widely accepted method, as it is one of the least expensive and technically demanding methodologies available

(Graves *et al.*, 2007). There are a number of excellent review papers detailing each of these methods and their applications to MST (Field and Samadpour, 2007; Meays *et al.*, 2004; Stoeckel and Harwood, 2007; Yan and Sadowsky, 2007).

1.1.2.2 Genotypic methods

Genotypic methods distinguish between sources by identifying patterns in the genetic material, or 'fingerprints' of bacterial isolates (Griffith *et al.*, 2003). Methods include pulse-field gel electrophoresis (PFGE), ribotyping, repetitive element sequencing-based PCR (Rep-PCR), amplified fragment length polymorphism (AFLP) and denaturing gradient gel electrophoresis (DGGE). PFGE is currently considered by many to be the 'gold standard' of molecular typing, as it is extremely sensitive to small genetic differences and results in high discrimination between isolates (Scott *et al.*, 2002; Simpson *et al.*, 2002). However, it has been suggested that this technique might be too sensitive to discriminate between host sources for MST studies, resulting in profiles that are too diverse for comparison (Lu *et al.*, 2004; McLellan *et al.*, 2003; Scott *et al.*, 2002). The review papers in section 1.1.2.1 also provide detailed information on the genotypic methods.

1.1.3 Library-independent methods

Over the past decade, LDMs have largely been replaced by LIMs, which do not require a large database of organisms and largely overcome the geographical diversity. Instead, LIMs rely on the detection of a particular host-specific organism or gene (Hagedorn *et al.*, 2011a). These methods differentiate between sources by identifying the presence of genetic markers that are unique to a particular faecal bacteria or targeted host species, therefore, operating at the population level rather than the isolate level (Field and Samadpour, 2007; Griffith *et al.*, 2003). This methodology also allows for the inclusion of many microorganisms which cannot be cultured using standard techniques, estimated at 99% of all microorganisms (Amann *et al.*, 1995; Su *et al.*, 2012). Anaerobic populations, such as *Bacteroidales*, are only cultivable for a couple of days after their introduction into extrainstestinal environments, such as surface water and groundwater (Bonjoch *et al.*, 2009; Okabe and Shimazu, 2007; Saunders *et al.*, 2009). In contrast, PCR amplifiable nucleic acid from *Bacteroidales* can persist for weeks in water (Dick *et al.*, 2010; Okabe and Shimazu, 2007; Walters and Field, 2009). Because of their limited

survival period in the environment, the presence of anaerobic bacteria in environmental samples represents recent faecal contamination events (Simpson *et al.*, 2002; Sinton *et al.*, 1998). LIMs have also been found to often show better accuracy in proficiency testing compared to LDMs (Griffith *et al.*, 2003; Harwood *et al.*, 2003; Myoda *et al.*, 2003).

"Ideal" requirements for a MST target have often been discussed, and some basic requirements have been suggested for source specificity and sensitivity. The ideal bacterial MST targets should only be present in the faecal material of the respective host source group being considered (specificity), and should be present in comparable numbers in the faeces of all subgroups of the targeted source, comparable to or exceeding concentrations of traditional faecal indicators (sensitivity) (Wuertz *et al.*, 2011). Other considerations for application include temporal, geographic and matrix applicability; repeatability; practicality and robustness, and time constraints (Hagedorn and Liang, 2011; Stoeckel and Harwood, 2007).

There are a number of bacterial species which have been explored for use in MST target-specific studies, including *E. coli* (Hamilton *et al.*, 2006; Lee, 2011; Ram *et al.*, 2007; Shanks *et al.*, 2007); *Bacteroidales* (Bernhard and Field, 2000a; Bernhard and Field, 2000b; Dick *et al.*, 2005a; Dick *et al.*, 2005b; Griffith *et al.*, 2003; Kildare *et al.*, 2007; Lee and Lee, 2010; Okabe *et al.*, 2007); *Bifidobacterium* (Bonjoch *et al.*, 2009; Hill *et al.*, 2010; King *et al.*, 2007; Lamendella *et al.*, 2008; Matsuki *et al.*, 2004); *Enterococcus* (Ahmed *et al.*, 2008; Byappanahalli *et al.*, 2008; Harwood *et al.*, 2004; Scott *et al.*, 2005; Soule *et al.*, 2006); *Faecalibacterium* (Dick *et al.*, 2005a; Dick *et al.*, 2005b; Zheng *et al.*, 2009); *Rhodococcus* (Gilpin *et al.*, 2002; Savill *et al.*, 2001; Tajima *et al.*, 2001); *Catellicoccus* (Lu *et al.*, 2008) and *Streptococcus* (Tajima *et al.*, 2001).

The "target" of a LIM study can be based on a variety of levels of interaction between microbe and host, including detecting a source-specific bacterial gene or its product, targeting a host-associated bacterial population, or a whole bacterial community. Most current MST methods target specific bacterial populations (Wuertz *et al.*, 2011). The abundant intestinal bacterial communities appear to differ from those found in extrainterstinal habitats (Ley *et al.*, 2008b), and diversification of gut microbiota appears to be related to host phylogeny, which supports the hypothesis that there are source-specific bacterial lineages (Ley *et al.*, 2008a; Ley *et al.*, 2006).

Currently, most LIMs are restricted by inadequate sensitivity, inability to quantify source contributions (Field *et al.*, 2003; Noble *et al.*, 2003) and possible geographical limitations (Ahmed *et al.*, 2007; Hamilton *et al.*, 2006). Few, if any, current methods have proven to be consistently unique to a specific species, detectable in significant quantities in environmental waters, and/or geographically stable in different regions (Hagedorn and Liang, 2011).

1.1.3.1 PCR-based methods

The majority of LIMs utilise PCR techniques to directly amplify and analyse a targeted gene, and generally do not require cultivation, which can save time and expense. The main advantage of PCR techniques is the potential for tracking several different genes concurrently, thereby allowing for a greater level of certainty in source tracking and pathogen detection (Santo Domingo *et al.*, 2007; Simpson *et al.*, 2002). The ability to assess the entire population within an environmental sample via PCR amplification avoids sample size biases that are a large contributing factor for problems with LDMs (Field *et al.*, 2003). PCR assays can also be performed in a matter of hours, and have the potential of being sensitive, inexpensive, quantitative and amenable to automation (Santo Domingo *et al.*, 2007). A challenge for PCR based methods, however, is the nucleic acid extraction and recovery step, as many environmental samples contain various PCR inhibitors, such as humic acids, complex polysaccharides and inorganic ions (Abu Al-Soud and Rådström, 1998; Roslev and Bukh, 2011; Simpson *et al.*, 2002).

1.1.3.1.1 Host-specific markers

Host-specific methods distinguish members of bacterial gene sequences by detecting differences in the nucleotide arrangements in a target sequence (Bernhard and Field, 2000a). A species-specific sequence must be chosen, often through creating a clone library using universal primers, where the clone inserts are amplified and sequenced, with the target sequence found by aligning DNA sequences from a database, such as GenBank. A primer pair can then be designed to specifically amplify the target sequence (Roslev and Bukh, 2011). The majority of host-specific studies have focused on the 16S ribosomal RNA (rRNA) gene (see section 1.1.4).

Other methods used to identify suitable primers include length heterogeneity-PCR (LH-PCR) (Bernhard and Field, 2000a; Bernhard and Field, 2000b; Suzuki *et al.*, 1998); terminal restriction fragment length polymorphism (t-RFLP) (Dick *et al.*, 2005b;

Fogarty and Voytek, 2005; Jeong *et al.*, 2008); subtractive hybridisation (Dick *et al.*, 2005a; Green *et al.*, 2012; Hamilton *et al.*, 2006; Zheng *et al.*, 2009) and genome fragment enrichment (Lu *et al.*, 2007; Shanks *et al.*, 2006).

Most of the original species-specific assays were developed for conventional PCR applications (Bernhard and Field, 2000a; Bernhard and Field, 2000b; Dick *et al.*, 2005a; Dick *et al.*, 2005b), which only provide a presence/absence result for a target sequence, and the PCR products can only be detected at the end of the protocol. Recently, real-time quantitative PCR (qPCR) assays have been designed, which quantify the markers during amplification, providing a rapid, quantitative detection of many targets (Roslev and Bukh, 2011). This is achieved via detection of either a non-specific fluorescent reporter, such as SYBR Green (Jeong *et al.*, 2010; Matsuki *et al.*, 2004; Okabe *et al.*, 2007; Seurinck *et al.*, 2005; Silkie and Nelson, 2009), or a specific labelled probe, such as TaqMan assays (Converse *et al.*, 2009; Dick and Field, 2004; Fremaux *et al.*, 2010; Kildare *et al.*, 2007; Layton *et al.*, 2006; Mieszkin *et al.*, 2009; Mieszkin *et al.*, 2010; Reischer *et al.*, 2006; Reischer *et al.*, 2007; Savill *et al.*, 2001; Shanks *et al.*, 2008; Shanks *et al.*, 2009; Siefring *et al.*, 2008) and Scorpion probes (Stricker *et al.*, 2008). No qPCR assay for host associated markers is absolutely specific and sensitive for its intended target, which may lead to false positive and negative information associated with each assay (Santo Domingo *et al.*, 2007; Wang *et al.*, 2010). It has been suggested that TaqMan assays are the better choice for working with environmental samples, due to the requirement of both probe and primer specificity; whereas SYBR Green chemistry also results in detection of non-specific products and messenger RNA (mRNA) with high sequence identity (Kildare *et al.*, 2007).

While most MST studies using host-specific markers have looked at their source-specificity and cross-amplification in a range of faecal sources, an increasing number of studies have also explored geographical applicability of these studies around the globe, including Australia (Ahmed *et al.*, 2007), Austria (Reischer *et al.*, 2007), Belgium (Seurinck *et al.*, 2005), Canada (Fremaux *et al.*, 2009; Lu *et al.*, 2011), France (Mieszkin *et al.*, 2009), Japan (Okabe *et al.*, 2007), Kenya (Jenkins *et al.*, 2009), Korea (Jeong *et al.*, 2008), New Zealand (Gilpin *et al.*, 2003; Kirs *et al.*, 2011) and the United States (Kildare *et al.*, 2007; King *et al.*, 2007; Lamendella *et al.*, 2008; Lu *et al.*, 2008; Shibata *et al.*, 2010).

1.1.3.1.2 Bacteroidales as targets

Among the first efforts of LIM development was the report that certain *Bacteroides spp.* were frequently associated with human faeces but not with animal faeces (Kreader, 1995; Stoeckel and Harwood, 2007), and some of the first host-specific MST assays used primers targeting the 16S rRNA of this genera (Bernhard and Field, 2000a; Bernhard and Field, 2000b). PCR assays targeting anaerobic members of the order *Bacteroidales* are currently the most widely used faecal source identifiers in water (Wuertz *et al.*, 2011), and have been shown to have more promise as source markers than a number of other species (Bernhard and Field, 2000a; Savichtcheva *et al.*, 2007). *Bacteroidales* constitute approximately one third of the human faecal bacterial population (Holdeman *et al.*, 1976), and are abundant in the digestive tract of other warm-blooded animals, although are generally not as dominant as in humans (Kreader, 1995; Meays *et al.*, 2004). Their abundance in avian species, including chicken, gull, turkey and goose, has been consistently reported as low (Fogarty and Voytek, 2005; Jeter *et al.*, 2009; Lu and Domingo, 2008).

There are currently four commonly used general assays that target all *Bacteroidales* - AllBac (Layton *et al.*, 2006), BacUni (Kildare *et al.*, 2007), GenBac (Dick and Field, 2004; Dick *et al.*, 2005a; Siefring *et al.*, 2008) and PreBac1 (Okabe *et al.*, 2007), as well as a panel of host-associated assays, including those specific for human, ruminant, dog, pig, horse, elk, gull, and Canada geese (Bernhard and Field, 2000a; Bernhard and Field, 2000b; Converse *et al.*, 2009; Dick *et al.*, 2005a; Dick *et al.*, 2005b; Dorai-Raj *et al.*, 2009; Fremaux *et al.*, 2010; Jeong *et al.*, 2010; Jeter *et al.*, 2009; Kildare *et al.*, 2007; Layton *et al.*, 2006; Mieszkin *et al.*, 2009; Mieszkin *et al.*, 2010; Okabe *et al.*, 2007; Reischer *et al.*, 2007; Reischer *et al.*, 2006; Seurinck *et al.*, 2005; Shanks *et al.*, 2008; Shanks *et al.*, 2009; Silkie and Nelson, 2009; Stricker *et al.*, 2008). This wide range of host-associated assays is advantageous when there are multiple faecal source contamination events in environmental samples, as it allows for multiplex PCR protocols to be used to detect and compare all potential sources at once (Wuertz *et al.*, 2011). However, there is always some cross-reactivity for all of these host-associated assays, as none are 100% specific for their host targets (McLain *et al.*, 2009; Silkie and Nelson, 2009). Of particular note is that there is currently no host specific assay for sheep. There are a number of assays which target cow-specific *Bacteroidales* sequences; however, many of these amplify all ruminant sources.

1.1.3.2 Alternative methods

1.1.3.2.1 Alternative gene markers

Alternative markers to the 16S rRNA gene studied have included a range of proteins, such as Enterococcal surface protein (Ahmed *et al.*, 2008; Byappanahalli *et al.*, 2008; Scott *et al.*, 2005), and host-specific protein targets from faecal anaerobes, many of which are unknown (Shanks *et al.*, 2007; Shanks *et al.*, 2006). Other gene markers have included mercury resistance genes (Bruce, 1997); the RNA polymerase β subunit (Case *et al.*, 2007); a range of *E. coli* genes (Field *et al.*, 2003; Khatib *et al.*, 2003; Lee, 2011; Savichtcheva *et al.*, 2007); *Enterococcus ddl* gene (Harwood *et al.*, 2004); *Bifidobacteria cpn60* gene (Hill *et al.*, 2010); human specific *Methanobrevibacter smithii nifH* gene (Johnston *et al.*, 2010); bacterial topoisomerase *gyrB* gene (Lee and Lee, 2010) and the human specific *Bacteroides thetaiotaomicron* α-1-6, mannanase gene (Yampara-Iquise *et al.*, 2008).

1.1.3.2.2 Virus targets

The direct measurement of human or bacterial viruses has been proposed as an alternative method for detecting faecal contamination, mainly due to the lack of correlation between faecal indicator bacteria concentrations with pathogen density and risk of gastrointestinal illness (Griffith *et al.*, 2003; McQuaig and Noble, 2011). Viruses studied include human enteroviruses (Fong *et al.*, 2005; Griffith *et al.*, 2003; Noble *et al.*, 2003); bovine enteroviruses (Fong *et al.*, 2005; Ley *et al.*, 2002), human adenoviruses (Fong *et al.*, 2005; Griffith *et al.*, 2003; Noble *et al.*, 2003); human polyomavirus (McQuaig *et al.*, 2006), bovine polyomavirus (Hundesa *et al.*, 2010) and F+ coliphages (Griffith *et al.*, 2003; Noble *et al.*, 2003).

Enteric viruses rely on specific cell surface receptors to bind to host cells, and have the advantage of inherent species specificity (Harwood, 2007). They are thought to be the causative agent of a large proportion of waterborne disease, so provide a more direct estimation of pathogen risk (McQuaig and Noble, 2011). Their high host-specificity, stability in different environments and prevalence in diverse geographical areas suggest they are a promising tool for MST methods (Hundesa *et al.*, 2010).

The use of RNA coliphages, particularly F+ RNA coliphages, which infect *E. coli*, is based on the observations that different serotypes are present in human and animal faeces, with types II and III generally associated with human faecal contamination, type

IV with animal and type I with both (Noble *et al.*, 2003). However, coliphages are usually identified in low numbers in environmental matrices and subject to temperature variation, thus confounding the use of direct assay and requiring enrichment procedures (Yan and Sadowsky, 2007).

1.1.3.2.3 Eukaryotic markers

Use of eukaryotic markers such as mitochondrial DNA (mtDNA) appears promising, as mitochondria are found in multiple numbers in all cells of eukaryotes. As mitochondria possess their own genome, they contain species-specific sequences due to their faster evolution compared with nuclear DNA (Caldwell *et al.*, 2011; Martellini *et al.*, 2005). Targeting mtDNA also allows for the faecal source organism to be identified directly, instead of microorganisms it might host (Roslev and Bukh, 2011). The idea of using mtDNA in MST was first proposed by Martellini *et al.* (2005), based on the fact that faeces contain larger amounts of cells from the host, such as epithelial cells from the intestines, and that these cells are excreted in the environment (Roslev and Bukh, 2011). Primers were designed for differentiating between human, bovine, porcine and ovine sources in surface water tracking; however, there were many problems with specificity and limits of detection (Field and Samadpour, 2007; Martellini *et al.*, 2005). More recent studies have shown that it is possible to design primers that are species-specific; however, mtDNA is shed by other means besides faeces, so detection of mtDNA from a particular organism may or may not indicate direct faecal contamination for this source (Roslev and Bukh, 2011).

1.1.3.2.4 Chemical methods

Chemicals that are specific to human wastewater offer several potential advantages over biological methods, as they typically require less sample preparation and analysis time, are generally unique to human-origin pollution, are not confounded by regrowth in the environment, and are more likely to be geographically and temporally stable (Hagedorn and Weisberg, 2009; Hagedorn *et al.*, 2011b). A large United States study looked for 110 different chemicals in wastewater samples, of which 78 were found at least once, and 35 suggested as potential human-specific indicators (Glassmeyer *et al.*, 2005); however, there is currently no chemical that has emerged as the best to use for human-specific contamination. Chemicals isolated in MST studies include pharmaceuticals (Hilton and Thomas, 2003; Roberts and Thomas, 2006), faecal sterols and stanols (Gregor *et al.*, 2002; Leeming and Nichols, 1996), fluorescent whitening agents

(Dickerson Jr *et al.*, 2007; Gilpin *et al.*, 2002) and caffeine (Chen *et al.*, 2002; Peeler *et al.*, 2006).

1.1.3.2.5 Microarrays

Microarrays consist of multiple DNA probes arrayed onto a solid surface to allow for fast, parallel, high-throughput multi-species detection (Bodrossy and Sessitsch, 2004; DeSantis *et al.*, 2007). The use of microarrays which target a range of phylogenetic markers and functional genes has been used to study microbial community structures, including a range of MST studies targeting the 16S rRNA gene (DeSantis *et al.*, 2007; Dubinsky *et al.*, 2012). Soule *et al.* (2006) screened sequences from a range of different host animals using a microarray composed of cloned DNA fragments from *Enterococcus*. There are potential advantages and limitations to using a phylogenetic microarray for source identification. An advantage is sensitive detection of taxa with low abundance in the community (DeSantis *et al.*, 2007; Dubinsky *et al.*, 2012). It also allows direct detection of pathogens which can be present in lower levels compared with indicators. However, probe design can limit detection to what is already known (Cardenas and Tiedje, 2008).

1.1.4 16S ribosomal RNA gene

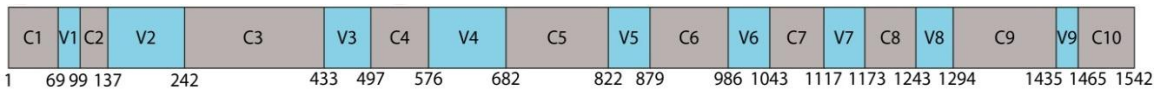
The 16S rRNA gene is currently the most commonly used species proxy for microbial community studies (Cardenas and Tiedje, 2008; Clarridge III, 2004). It is universal in bacteria, with its sequence approximately 1,550 bp long, and is composed of conserved regions interspaced with nine hypervariable regions (V1-V9) (Chakravorty *et al.*, 2007) (Figure 1.1). The conserved regions have a low rate of evolution and therefore have a very similar sequence across all bacterial species, while the hypervariable regions demonstrate considerable sequence diversity among different bacteria species, constituting useful targets for taxonomic identification between organisms at the genus level across all phyla of bacteria. There are a number of dedicated rRNA databases available, including the Ribosomal Database Project (Cole *et al.*, 2009), SILVA (Quast *et al.*, 2013), Greengenes (DeSantis *et al.*, 2006), and EzTaxon-e (Kim *et al.*, 2012), which can all be used to classify partial and full-length 16S rRNA sequences. There are also multiple copies of the 16S rRNA gene in each bacterial cell, which increases sensitivity, aiding detection in environmental samples (Dick and Field, 2004; Shanks *et*

al., 2008). However, these intragenomic copies can differ in sequence, which can result in identification issues (Case *et al.*, 2007).

A large number of primer sequences for amplification and sequencing of rRNA genes have been published, some designed as taxa specific, whilst others have been designed to amplify all prokaryotic genes, referred to as 'universal' primers (Baker *et al.*, 2003). These have also targeted a range of hypervariable regions, with differing results; however, the majority of these studies have focused on the V1, V2, V3, V4, or V6 regions (Figure 1.1). It is generally accepted that one hypervariable region does not provide enough diversity to identify different species (Chakravorty *et al.*, 2007), so most studies include at least two hypervariable regions in their target sequence.

Although the use of the 16S rRNA gene has known limitations, including low rate of evolution, lack of correlation with organism function and variable copy number, no other molecular marker has emerged that is found in all bacteria, has as low a rate of horizontal gene transfer and recombination, or has sufficient genetic information to differentiate closely related organisms (Schloss *et al.*, 2011).

Figure 1.1: Conserved and hypervariable regions of the 16S rRNA gene. The conserved regions (C1-C10) are shown in grey and the hypervariable regions (V1-V9) in blue. Nucleotide positions are based on E. coli nomenclature.



1.2 DNA Sequencing

The ability to sequence DNA has revolutionised the way molecular biology is undertaken. Knowing the sequence of a target provides the ability to gain important information on genes, genetic variation and gene function, rapidly advancing our understanding of the environment, health and disease. For many years, the largest limitation in DNA sequencing was cost, estimated at $500 - $1,500 per Mb (Glenn, 2011; Kircher and Kelso, 2010), making many genomic projects unachievable; however, recent advances in technology have opened up possibilities to rapidly generate

large-scale sequencing data at a reasonable cost, and we are now entering an era where almost any organism can ‘go genomic’ (Ekblom and Galindo, 2011).

1.2.1 First generation sequencing

The first genome to be completely sequenced was the bacteriophage ϕX174, approximately 5,375 nt in size, in 1977 (Sanger *et al.*, 1977a). This utilised a new technology, termed Sanger sequencing, based on chain-termination chemistry and taking advantage of the ability of DNA polymerase to incorporate dideoxy nucleotide triphosphates (ddNTP), nucleotide analogues that lack the 3’-hydroxyl group essential in DNA bond formation (Sanger and Coulson, 1975; Sanger *et al.*, 1977b). The original method was primarily manual, and utilised isotopic radioactive labelling of primers for imaging, using a ‘plus and minus’ method (Sanger and Coulson, 1975). Since this first publication, the sequencing of genomes, individual regions and genes has become a major focus of modern biology, and current Sanger sequencing systems, such as the Applied Biosystems 3xxx series or the GE Healthcare MegaBACE instrument, are still based on the same general scheme applied in 1977, and are considered ‘first generation’ sequencing technology. The system has, however, undergone a steady metamorphosis over the years, including the radioactive labelling replaced with four differently coloured fluorophore tags on each of the ddNTPS, capillary electrophoresis replacing the use of slab-gel electrophoresis for separation of fragments, and parallelisation of sequencing runs, resulting in a large-scale production that is almost completely automated (Hutchison III, 2007). These developments allowed the system to be used to sequence the first human genome, a project that took a decade and cost $US3 billion to complete (Collins *et al.*, 2004; Venter *et al.*, 2001).

Multiple copies of the target sequence are cloned, purified and amplified, followed by reverse strand synthesis using a known priming sequence upstream of the target sequence. Each sequencing reaction requires a mixture of deoxy nucleotide triphosphates (dNTPs) and a fluorescently labelled ddNTP. DNA polymerase adds either a dNTP or the corresponding ddNTP at each step of chain extension, with the ddNTP incorporation causing random, non-reversible termination of the synthesis reaction. Amplification results in multiple fragments of the same molecule extended to different lengths, which can be resolved based on size using capillary electrophoresis,

where the signal from the terminating fluorescently labelled base determines the sequence. Current Sanger sequencing technology can resolve up to 384 sequences simultaneously (Emrich *et al.*, 2002), between 600 and 1,000 nt in length (Hert *et al.*, 2008; Kircher and Kelso, 2010), although standard 96-capillary instruments generally yield an average of 800 bases (Schadt *et al.*, 2010). This results in approximately 6 Mb of DNA sequence per day, with costs amounting to $500 - $1,500 per Mb (Glenn, 2011; Kircher and Kelso, 2010). While still frequently used, the major limitations of this method include the low throughput, with only small amounts of DNA able to be processed per unit of time, as well as the high cost (Schadt *et al.*, 2010). The reliance on electrophoresis for separation has ultimately led to this technology reaching its pinnacle for both speed and cost, with little ability left to further increase speed or a higher degree of parallelisation, leading to the requirement of a completely new generation of approaches to DNA sequencing (Zhou *et al.*, 2010a).

1.2.2 Second generation sequencing

Over the past ten years, alternative sequencing strategies have emerged, resulting in much higher throughput sequencing, which can outperform Sanger sequencing technologies up to a factor of 1,000 times in daily throughput. This leads to a reduction in the cost, with sequencing one million nucleotides costing less than 1% of that associated with Sanger sequencing (Kircher and Kelso, 2010). The underlying principle of next generation sequencing involves the DNA molecules, which are sequenced millions at a time, in a massively parallel fashion. This can be achieved in a stepwise repetitive process, or in a continuous real-time manner, where each individual template is independently sequenced and counted among the total sequences generated (Pareek *et al.*, 2011).

The majority of DNA sequencing studies are now performed using second generation sequencing, or next generation sequencing (NGS) instruments, of which two currently dominate the commercial market: the Roche 454 Genome Sequencer (GS) and the Illumina Genome Analyzer. Both companies now also offer alternative systems, such as the GS 454 Junior and the Illumina MiSeq, which are considered benchtop versions, and the Illumina HiSeq. Other platforms available include SOLiD (Applied Biosystems) and Ion Torrent (Life Technologies). All these technologies are based on various strategies

that rely on a combination of template preparation, clonal amplification, sequencing and imaging, as well as genome alignment and assembly methods (Metzker, 2010). These distinctions in strategies, however, have led to a large variability among and within the NGS platforms in terms of template size and construct, read-length, throughput and coverage, with each platform generating its own pattern of bias (Harismendy *et al.*, 2009; Metzker, 2010).

1.2.2.1 Sequencing providers

Due to the rapid advancement of the technology in this field, and the high cost associated with purchasing and maintaining these sequencing instruments, it is not feasible for all laboratories to operate their own. Instead, many laboratories make use of the many sequencing providers around the world who offer a range of sequencing technologies and data analysis options at competitive prices. Currently only one company offers these services in New Zealand, New Zealand Genomics Limited (NZGL), who work in collaboration with many of the universities who own and operate a range of next generation sequencing platforms. There are also many global sequencing companies who offer a range of sequencing services, including Macrogen Inc. (Korea), BGI (China) and CD Genomics (USA), as well as many universities who offer services to external parties.

The two key platforms offered by these services are the Roche 454 GS FLX and the Illumina platforms. An in-depth review on how these two systems work is detailed below, as well as a general introduction to the SOLiD and Ion Torrent systems. General comparisons of the performance of the different systems are provided in Table 1.1, however, it must be noted that the technology behind each of these platforms is improving constantly, and therefore the values provided may not be entirely accurate. It must also be noted that there is currently no accepted standards for what measures companies need to report, or how data is analysed, which can have significant impact on how the platforms compare (Glenn, 2011).

1.2.2.2 Roche 454 GS FLX

The 454 sequencing platform was the first of the NGS platforms on the market (Margulies *et al.*, 2005), and utilises the pyrosequencing approach developed by Nyren and Ronaghi (Ronaghi *et al.*, 1996). This approach, often called "sequencing by synthesis", is based on the detection of light emitted when a single nucleotide is

incorporated by a DNA polymerase, so does not require a physical separation process like electrophoresis to resolve the next base in the DNA strand, offering advantages of real-time detection.

1.2.2.2.1 Library preparation

Next generation sequencing is currently limited to short fragments of DNA, so samples must be sheared into smaller fragments prior to sequencing. For high molecular weight DNA samples, such as genomic DNA, this involves nebulisation, which physically shears dsDNA into fragments ranging from 400-1,000 bp. This generates a library of small sized DNA fragments, produced from a single DNA sample. Following fragmentation, specific double stranded "adaptor" oligonucleotides, termed Adaptors A and B, are ligated to the ends of each DNA fragment, which provide priming regions to support amplification and nucleotide sequencing, as well as a four-base sequencing "key" sequence. Adaptor B also contains a biotin tag on its 5' strand, which allows for immobilisation of the dsDNA fragments onto magnetic streptavidin-coated beads. The dsDNA is denatured to release the complementary non-biotinylated strands, forming a ssDNA template library of fragments which contain both the A and B adaptor sequences.

Once a single-stranded DNA library is constructed, emulsion-based clonal amplification (emPCR), initially described by Dressman *et al.* (2003), is utilised to produce thousands of copies of each library fragment in a cell-free system. This avoids the arbitrary loss of genomic sequences which is inherent in bacterial cloning methods, such as those used in Sanger sequencing (Metzker, 2010); however, it instead may introduce PCR-based biases (Harris *et al.*, 2008). The ssDNA fragments are immobilised by hybridisation onto a second set of beads, DNA Capture beads, containing the complementary Adaptor A and B sequences. The system has been optimised so that each bead will only bind to a single fragment (Margulies *et al.*, 2005). The beads are then emulsified with amplification reagents in a water-in-oil mixture, which traps individual beads in their own microreactor, allowing the clonal amplification to proceed, generating millions of copies of each template strand.

For samples that have a low molecular weight, such as short PCR products and DNA derived from microRNAs, amplicon libraries can be generated directly, without the need for nebulisation or ligation of adaptors. Instead, the procedure consists of a PCR

amplification using fusion primers, which contain the required sequencing Adaptors A and B, the sequencing key, and a template specific region.

1.2.2.2.2 Sequencing

Once amplification is completed, the microreactors are broken, and the DNA Capture beads containing the amplified fragments are layered onto a fibre-optic slide, known as a PicoTiterPlate (PTP) device, which is a 75 mm x 75 mm support of 3.5 million optical fibres etched with wells (Leamon *et al.*, 2003; Margulies *et al.*, 2005; Rothberg and Leamon, 2008). The DNA Capture beads are surrounded by smaller enzyme beads and packing beads; the enzyme beads have ATP sulfurylase and luciferase attached to them, which are key components of the sequencing reaction, while the packing beads ensure the DNA beads remain positioned in the wells. All of the required reagents are flowed across the wells of the plate, with nucleotides introduced sequentially. DNA polymerases incorporate each complementary nucleotide as it is washed across the template strands, and incorporation pauses once the longest possible stretch of complementary nucleotides has been synthesised. In the process of incorporation of each nucleotide, a pyrophosphate moiety (PPi) is released, which is then converted to adenosine triphosphate (ATP) by the enzyme ATP sulfurylase. The ATP is hydrolysed by luciferase, which converts ATP and luciferin into oxyluciferin, emitting a light signal (Ronaghi *et al.*, 1996; Ronaghi *et al.*, 1998). The light emission is detected for all wells in parallel by a high resolution charge-coupled device (CCD) camera, where the intensity of the light from each individual well is proportional to the number of nucleotides incorporated. Since ATP is a substrate of luciferase for the sequencing reaction, deoxy-adenosine-5' (α-thio)-trisphosphate, which is not a substrate of luciferase, is provided for incorporation in the DNA strand reaction (Kircher and Kelso, 2010). By-products and unincorporated reagents are then washed away, and the next nucleotide in the predetermined sequence is flowed across the PTP device.

1.2.2.2.3 Imaging and data processing

The incorporated data processing software uses the signal intensity from each incorporation event at each well position to determine the sequence of all the reads in parallel. The current 454 GS FLX Titanium XLR70 platform can sequence up to one million reads in a single experiment, determining sequencing up to 600 bp in length with an average read length of 450 bp, a typical throughput of 450 Mb and 99.9% consensus accuracy. A recent development of this system, the GS FLX Titanium XL+

can now produce read lengths of up to 1,000 bp, with an average of 700 bp, although it requires a longer run time of 23 h, compared with 10 h for the XLR70 system, and is currently not suited for all applications, such as amplicon sequencing.

1.2.2.2.4 Advantages and limitations

A major limitation of the 454 technology is the resolution of homopolymer-containing DNA segments (Wicker *et al.*, 2006; Zhou *et al.*, 2010b), due to the use of light emitted to determine the number of repetitive bases. There is a higher error rate associated with the determination of the exact number of bases compared to the discrimination of incorporation verses non-incorporation, therefore, the dominant error type is insertions and deletions, rather than substitutions (Quinlan *et al.*, 2008; Zhou *et al.*, 2010b). However, higher coverage and computational post-processing may correct for many of these issues (Green *et al.*, 2006; Kircher and Kelso, 2010). Phasing is another problem associated with 454 sequencing, where a population of DNA molecules amplified from the same starting molecule do not all get extended properly in every cycle. This results in a loss of synchronicity, or dephasing, and causes an increase in noise and sequencing errors as the read extends (Erlich *et al.*, 2008; Schadt *et al.*, 2010).

The key advantage of the 454 platform over other NGS methods is its read length, which makes it ideal for metagenomics projects, where the more information gathered about a sequence, the higher the probability of identifying which organism it is from (Zhou *et al.*, 2010b). Longer reads also allow single direction sequencing to be performed, reducing the amount of computational assembly normally required for paired-end reads where sequencing occurs from both directions. Unlike many of the other second generation NGS platforms, 454 pyrosequencing does not need to carry out a chemical deblocking step to allow DNA extension to continue, which reduces the likelihood of premature chain termination and non-simultaneous extension, limiting the effects of dephasing (Metzker, 2010; Zhou *et al.*, 2010a). The 454 PTP is also set up to allow eight individual sequencing reactions to occur simultaneously, allowing scalability dependent on the project. A half-sized plate region will produce approximately 500,000 reads, with an eighth region plate resulting in approximately 100,000 reads. The GS Junior platform also results in approximately 100,000 reads, although it only offers one reaction plate per run.

1.2.2.3 Illumina Genome Analyser/HiSeq

The Illumina platform is also based on sequencing by synthesis, but utilises reversible nucleotide terminator technology, where the incorporation reaction is stopped after each base. This utilises a similar methodology to Sanger sequencing; however, this approach generates several billion reads of nucleotide sequence (Bentley *et al.*, 2008). The initial "Solexa-sequencing" was introduced in 2006, and allowed for the simultaneous sequencing of several million, very short sequences, which were less than 26 nt long. Continuous developments by Illumina now allow sequencing of up to 500 bp through paired-end sequencing on their MiSeq platform.

1.2.2.3.1 Library preparation

As for the 454 sequencing approach, the Illumina technology requires DNA samples to be converted into a special library (Bentley *et al.*, 2008), the input DNA needs to be fragmented into short lengths and two end-specific adaptors ligated to the ends of each fragment. The prepared library is denatured and hybridised to a "lawn" of oligonucleotides immobilised to a glass surface, or flow cell (Kircher and Kelso, 2010). The adaptors act as primers for the following "bridge" amplification (Adessi *et al.*, 2000; Fedurco *et al.*, 2006), where reverse strand synthesis starts from the hybridised, double stranded section of the template. As each new strand is synthesised, it can bend over and attach to another oligonucleotide bound to the flow cell, which is complementary to the second adaptor sequence. This results in the synthesis of the second covalently bound reverse strand, producing "bridges" of sequences bound at both ends to the flow cell. This amplification process is repeated several times to produce randomly distributed, clonally amplified clusters of approximately 1,000 copies of the original sequence, and the solid-phase amplification step producing up to 200 million separated clusters (Metzker, 2010). The dsDNA is denatured to obtain a ssDNA library, to ensure the sequencing reaction is not hindered sterically or by complementary base pairing (Kircher and Kelso, 2010).

1.2.2.3.2 Sequencing

Illumina platforms utilise cyclic reversible termination chemistry, using a set of four reversible terminators which are each labelled with a different removable fluorophore (Turcatti *et al.*, 2008). Sequencing is initiated by the hybridisation of a primer complementary to the adaptor sequence, which is followed by addition of polymerase and a mixture of the four terminators (Voelkerding *et al.*, 2009). Simultaneous addition

of all four terminators ensures incorporation in a step-wise manner, driving incorporation to completion with no risk of over-incorporation, while also minimising the risk of mis-incorporations (Bentley *et al.*, 2008). After incorporation, the remaining unincorporated nucleotides are washed away, and imaging using a CCD camera occurs to detect the incorporated nucleotide and position. This is followed by a cleave step, utilising tris (2-carboxyethyl) phosphine (TCEP) to remove the 3' terminating group and dye, regenerating the 3' hydroxyl group for the next cycle of nucleotide addition (Bentley *et al.*, 2008).

Originally, the Illumina platform could only work from one direction; however, an upgrade in 2008 resulted in the ability to complete a second round of synthesis from the opposite end of each strand (Zhou *et al.*, 2010b). After the initial rounds of synthesis, the newly sequenced strands are stripped off through chemical melting and washing, and the bridge amplification is repeated for a couple of cycles for reverse strand synthesis. The starting strand is selectively removed before annealing another sequencing primer for the second read, where a full set of cycles of bridge amplification is repeated, forming a new set of clusters (Kircher and Kelso, 2010). This "paired-end" approach enables up to twice the amount of data to be generated, and can effectively double the length of sequence able to be determined.

1.2.2.3.3 Imaging and data processing

After incorporation, an imaging step follows, during which the flow cell is imaged in three 100-tile segments by the CCD camera (Mardis, 2008), where the unique fluorophore for each terminator reveals the identity of the newly incorporated nucleotide for each cluster (Zhou *et al.*, 2010a). This is achieved through the use of total internal reflection fluorescence, via the use of red and green lasers (Kircher *et al.*, 2009; Metzker, 2010). After sequencing, the images are analysed and the intensities for each cluster are extracted. A base-calling algorithm assigns sequences and associated quality values to each read (Zhou *et al.*, 2010b). The Illumina HiSeq 2500 is the largest of the Illumina platforms, generating up to 600 GB of data, and 6 billion paired-end reads, although is currently limited to a read length of 2 x 100 bp, and can take 10 days to complete a high-output run. The Illumina MiSeq is designed as a benchtop sequencer, and although it has a much smaller throughput of only 8 GB, and 34 million paired-end reads, it is now able to sequence read lengths of up to 500 bp, through paired-end reads

of 250 bp. It is also much faster, taking approximately 39 h to complete a 2 x 250 bp sequencing run.

1.2.2.3.4 Advantages and limitations

The Illumina library and flow cell preparation includes several *in vitro* amplification steps, which results in a high background error rate (Kircher and Kelso, 2010), contributing to an average error rate of 10^{-2} – 10^{-3} (Dohm *et al.*, 2008; Kircher *et al.*, 2009), with substitutions the most common error type (Metzker, 2010). As is the case for the other next generation sequencing systems, the error rate increases with increasing length of determined sequence, mainly due to phasing and fluorophore intensities declining over time (Erlich *et al.*, 2008; Kircher *et al.*, 2009). Phasing occurs when nucleotides are under- or over-incorporated in a given sequencing cycle, resulting in a cluster producing a heterogeneous population of strands of varying lengths (Voelkerding *et al.*, 2009). Simultaneous identification of all four fluorophores can also be an issue, as two pairs (A/C and G/T) are excited using the same laser, present similar emission spectra and show only limited separation using optical filters (Kircher and Kelso, 2010).

Illumina's biggest advantage is its cost per Mb of sequence, at approximately $0.10/Mb, compared to around $10/Mb for 454 sequencing platforms (Glenn, 2011), but the shorter sequencing lengths require the use of paired-end sequencing to produce reads long enough for taxonomic identification, which needs assembly tools to map the reads together.

1.2.2.4 Alternative sequencing platforms

1.2.2.4.1 Applied Biosystems SOLiD

The Support Oligonucleotide Ligation Detection system (SOLiD) was initially published in 2005 (Shendure *et al.*, 2005), and the technology was made commercial in late 2007, making the SOLiD platform the third NGS system on the market (Kircher and Kelso, 2010). SOLiD utilises sequencing by ligation chemistry, which is based on sequential ligation with DNA ligase, rather than DNA polymerase, using a set of fluorescently labelled hybridisation probes, which can be ligated to specific primers (Housby and Southern, 1998; Shendure *et al.*, 2005; Zhou *et al.*, 2010b).

As with the other systems, only small fragments can be sequenced, so template DNA must be fragmented to construct a suitable library. The DNA fragments are ligated to specific adaptor sequences and undergo emPCR, similar to the system used by 454 sequencing (Kircher and Kelso, 2010; Pareek *et al.*, 2011). After library amplification, the templates are modified at the 3' end and covalently bound to a glass slide, creating a random dispersion of beads in a sequencing chamber, with at least 300 million beads per slide (Hert *et al.*, 2008; Kircher and Kelso, 2010; Zhou *et al.*, 2010b). During the sequencing reaction, a mixture of four fluorescently labelled octamers is added, which compete for ligation to the sequencing primer. The octamers comprise of three degenerative bases, three universal bases and two interrogation bases (Hert *et al.*, 2008; Metzker, 2010). With detection of the fluorescent label, the first two bases of the template sequence are determined and the ligated oligonucleotide probe is cleaved after the 5^{th} base, which leaves a free 5' phosphate on the extended primer for the following round of ligation (Kircher and Kelso, 2010). The ligation cycle is repeated for seven cycles, each cycle determining bases 6 and 7 of the template sequence. The template is then denatured, removing the ligation product, allowing the template strand to be reset with another octamer, for a second round of ligation cycles (Metzker, 2010; Zhou *et al.*, 2010b). This results in a "two-base-encoding" system, where each base has been interrogated in two independent ligation reactions by two different primers, resulting in a very powerful discrimination technique (Zhou *et al.*, 2010b).

Because the ligation reaction is based on probe recognition, rather than sequential addition, it is less prone to the accumulation of errors compared to the other NGS platforms. As with the Illumina platform, the random dispersion of the beads on the glass plate complicates identification of images, and results in the possibility of other objects, such as chemical crystals, dust and lint particles being misidentified as clusters (Kircher and Kelso, 2010).

1.2.2.4.2 Life Technologies Ion Torrent PGM

The Ion Torrent Personal Genome Machine (PGM) was launched in early 2011 (Rothberg *et al.*, 2011), and is based on the use of a disposable massively parallel semiconductor-sensing device, the Ion chip. It utilises a well characterised biochemical process, the release of a hydrogen ion as a by-product of DNA polymerase incorporating a nucleotide into a strand of DNA (Pareek *et al.*, 2011).

A fragmented DNA library is constructed, with ligation to specific adaptors, and clonally amplified using emPCR. The templates are then applied to the Ion chip, where the sequencing primers and DNA polymerase are bound to the template-carrying beads, and deposited into the chip wells (Rothberg *et al.*, 2011). During sequencing, all four nucleotides are provided in a stepwise fashion. When a complementary base contacts a template bead, the nucleotide is incorporated by the bound polymerase, and results in the hydrolysis of the incoming nucleotide. This causes a single proton to be released into solution for every nucleotide incorporated, resulting in a shift in the pH of the surrounding solution, proportional to the number of nucleotides incorporated (Moorthie *et al.*, 2011; Rothberg *et al.*, 2011). Beneath the wells is an ion-sensitive layer and a proprietary Ion sensor, which can detect the change in pH of the solution without scanning, cameras and light, resulting in real-time detection with no modified reagents required (Moorthie *et al.*, 2011; Pareek *et al.*, 2011). The change is converted to a voltage and is digitalised by off-chip electronics, all occurring within four seconds (Rothberg *et al.*, 2011). After the flow of each nucleotide, a wash is used to remove any unincorporated nucleotides and the system is run again, with the next nucleotide. If a nucleotide that floods the chip is not a match to the template strand, no voltage change will be recorded and no base will be called. If there are two identical bases next to each other on the strand, the voltage recorded will be doubled, and the chip will record two identical bases.

Ion Torrent offer a range of differently sized Ion chips, with increasing throughput for each, providing scalability and flexibility, due to the ability to choose the most suitably sized chip for the project. Generally, the smaller chips operate faster, but with lower throughput, with the smallest chip expected to take 2.4 hours to generate 20 Mb of 200 bp reads. The completion of a whole sequencing run in such a short time is a major advantage for the Ion Torrent, although currently the read lengths are similar to those produced by Illumina. A recent study also found the Ion Torrent to have the highest rate of insertion and deletion errors, averaging 1.72 errors per read, and was the least accurate compared to the Roche 454 GS Junior and the Illumina MiSeq when calling homopolymers (Loman *et al.*, 2012).

1.2.3 Third generation sequencing

A new generation of single-molecule sequencing is also emerging, based on sequencing from a single DNA molecule, without the need for a prior amplification step. This amplification has been suggested to be problematic for sequencing due to variable efficiencies as a function of template properties, introduction of uncontrolled bias in template representation, and introduction of errors (Harris *et al.*, 2008). Current single-molecule sequencing platforms on the market are the PacBio RS (Pacific Biosciences) and the Heliscope (Helicos BioSciences), but there are other technologies continuing to be developed, such as the GridION system (Oxford Nanopore Technologies), which utilises biological molecules engineered to form nanopores, with individual nucleotides cleaved off as they pass through (Clarke *et al.*, 2009; Howorka *et al.*, 2001; Stoddart *et al.*, 2009). General comparisons for the two currently available platforms are included in Table 1.1.

This third generation of sequencing (TGS) technologies promises advantages over current sequencing technologies in a number of ways: higher throughput, faster turnaround time, longer read lengths, higher consensus accuracy, small amounts of starting material and low cost (Pareek *et al.*, 2011; Schadt *et al.*, 2010). There are numerous platforms at different stages of development, and as these progress over the next coming years, third generation sequencing should once again change how we go about DNA sequencing.

1.2.3.1 Helicos BioSciences Heliscope

The Heliscope (Harris *et al.*, 2008) was the first single molecule sequencing platform on the market, launched in 2009, and is based on a similar methodology to that used for Illumina (Moorthie *et al.*, 2011). However, because it is a single molecule being sequenced, all the nucleotides need to be added individually, with sequencing halted to determine which nucleotide is incorporated (Bowers *et al.*, 2009; Schadt *et al.*, 2010). Nucleotides are fluorescently labelled and act as a terminator, allowing imaging to determine the identity of each nucleotide after incorporation. Chemical cleavage of the fluorophore allows progression to the next cycle with another fluorescently labelled nucleotide. The Heliscope is also capable of sequencing RNA directly, by replacing DNA polymerase with a reverse transcriptase enzyme (Ozsolak *et al.*, 2009), without requiring the conversion of RNA to complementary DNA (cDNA) or

ligation/amplification steps, as the second generation sequencing platforms require (Schadt *et al.*, 2010).

1.2.3.2 Pacific BioSciences PacBio RS

The PacBio RS platform uses single molecule real time (SMRT) sequencing, carried out on a sequencing chip containing thousands of zero-mode waveguides (ZMWs). A ZMW is a small hole in a metal film, deposited on a glass surface. Visible laser light cannot pass entirely through, and exponentially decays as it enters the ZMW, so by shining laser light up through the ZMW, only fluorescent labels inside the hole are excited, effectively eliminating background noise (Eid *et al.*, 2009; Levene *et al.*, 2003; Schadt *et al.*, 2010). A DNA polymerase molecule is attached to the bottom of each ZMW, and sequencing occurs as fluorescently labelled nucleotides are flooded across the array, travelling down into the ZMW and diffusing back out. As no laser light reaches the top of the holes to excite fluorescent labels, only the incorporated nucleotides are excited (Eid *et al.*, 2009; Schadt *et al.*, 2010). The difference between this method and others that use fluorophores is that the dye is attached to the phosphate of the nucleotide, and so is cleaved and released as a natural part of the synthesis reaction, allowing real time detection (Moorthie *et al.*, 2011).

1.2.4 Comparison studies

A number of comparison studies have been undertaken in recent years, comparing a range of the NGS platforms on offer (Archer *et al.*, 2012; Claesson *et al.*, 2010; Loman *et al.*, 2012; Quail *et al.*, 2012b; Ratan *et al.*, 2013; Suzuki *et al.*, 2011). All sequencers appear to complete the desired projects, with slightly varying abilities and errors reported, and the preferred sequencer for each study is often dependent on the sequencing project undertaken. All six comparisons reported here compared a 454 instrument and an Illumina instrument, with the exception of Quail *et al.* (2012b) who did not use a 454 instrument. The Ion Torrent was included in three (Archer *et al.*, 2012; Loman *et al.*, 2012; Quail *et al.*, 2012b), with PacBio RS (Archer *et al.*, 2012; Quail *et al.*, 2012b) and SOLiD (Ratan *et al.*, 2013; Suzuki *et al.*, 2011) each in two studies.

1.2.4.1 Human genome studies

Archer *et al*. (2012) looked at the V3 region of the HIV-1 *env* gene from 12 patient samples. All four platforms tested showed similar results for detection and sensitivity, suggesting that any of these NGS methods are suitable for predicting HIV-1 coreceptor usage. Most of the reported errors were comparable, with the Ion Torrent platform having the most number of deletion and insertion errors, and a close second to the 454 for substitutions. Illumina performed the best in terms of errors, with the lowest number of deletion and insertion events, while the PacBio had the lowest level of substitutions. Ratan *et al*. (2013) studied a portion of the human genome, comparing three platforms for their ability to identify single-molecule substitutions. Again, all platforms performed similarly, although the validation rate for variants supported by more than one platform were higher compared to the rate from individual platforms, suggesting the use of multiple platforms for assessing variants. Different factors that affect the ability to accurately call variants in the human genome were considered, with the unbiased distribution of reads across the genome considered one of the more important factors. The 454 platform was found to produce the most uniformly aligned data, despite having the lowest coverage.

1.2.4.2 Microbial genome studies

The other four studies all looked at microbial genomes, ranging from variable 16S regions (Claesson *et al.*, 2010) to the complete genome for four bacteria species (Quail *et al.*, 2012b), while the other two studies looked at different *E. coli* isolate genomes (Loman *et al.*, 2012; Suzuki *et al.*, 2011). Quail *et al*. (2012b) found the PacBio RS produced the highest error rates and least accurate data, and noted the high cost per base limited the large scale use of this platform. Suzuki *et al.* (2011), however, noted the Illumina GA resulted in the poorest accuracy and high error rates, with SOLiD giving the highest amount of 'junk' data, with only half the reads able to be aligned to the reference sequence. Loman *et al.* (2012) also sequenced a single *E. coli* genome, and noted that while all three benchtop platforms were able to generate a useful draft genome sequence, they did so quite differently, with "sequenced" meaning different things for different platforms. The Illumina MiSeq had the highest throughput and lowest error rates, the 454 GS Junior generated the longest read lengths but with the lowest throughput, and the Ion Torrent had the fastest run time, but the shortest reads and highest error rates. Claesson *et al*. (2010) sequenced various regions of the 16S

rRNA gene, mainly comparing how the different regions performed, but concluded that the 454 platform was currently the best option. This was predominantly due to the much longer read length produced by the 454 platform, however, it was noted that as Illumina developed longer read lengths it would potentially become a better option due to its higher output and lower costs.

1.2.5 Choosing the right platform

Each platform has its own strengths and weaknesses, which must be taken into account when choosing the most suitable platform for each project. From looking at these comparison studies, the Illumina platforms generally produce the most data at the lowest cost, but with mixed performances in regards to error rates. The longer length generated by the Roche 454 platforms give this technology advantages for assemblies of whole genome studies, as they result in less contigs overall (Suzuki *et al.*, 2011), and higher classification efficiencies for taxonomic assignments for microbial community studies (Claesson *et al.*, 2010). However, the 454 platforms have a much higher cost, generally an order of magnitude higher than for the other platforms (Kircher and Kelso, 2010; Loman *et al.*, 2012). The Ion Torrent is one of the newest platforms on the market, and has shown the greatest improvement in performance over a short period of time. The ability to scale the performance of the sequencing through the use of the disposable Ion chip suggests that this platform will be much more flexible than the others, but does appear to have quite high error rates. The SOLiD and PacBio systems, while able to generate the desired results for each study, do not appear to offer the best data available, with high error rates reported for both systems.

Table 1.1: Comparison of current sequencing instruments. Where possible, data were taken directly from information provided by the manufacturer; where this was not possible data were taken from a range of reviews (Glenn, 2011; Liu *et al.*, 2012; Metzker, 2010).

Instrument	Sequencing mechanism	Average Read length	Yield / run (Throughput)	Reads/run	Run time
Applied Biosystems 3730xl	Dideoxy chain termination	800 bp	0.06 Mb	96	2 h
Roche 454 GS FLX Titanium XLR70	Pyrosequencing	450 bp	450 Mb	1 million	10 h
Roche 454 GS Junior	Pyrosequencing	450 bp	35 Mb	100,000	10 h
Illumina GA IIx	Reversible terminator	2 x 150 bp	95 Gb	640 million (paired-end)	14 days
Illumina HiSeq	Reversible terminator	2 x 100 bp	600 Gb	6 billion (paired-end)	3 – 11 days
Illumina MiSeq	Reversible terminator	2 x 250 bp	7 Gb	30 million	35 h
Applied BioSystems SOLiD 5500xl	Sequencing by ligation	75 + 35 bp	120 Gb	1.4 billion	7 days
Life Technologies Ion Torrent PGM 314 chip	Semiconductor chip	35, 200 or 400 bp	3 – 40 Mb	100,000	0.5 – 3.7 h
Life Technologies Ion Torrent PGM 316 chip	Semiconductor chip	35, 200 or 400 bp	30 – 400 Mb	1 million	0.7 – 4.9 h
Life Technologies Ion Torrent PGM 318 chip	Semiconductor chip	35, 200 or 400 bp	0.3 – 1 Gb	5 million	0.9 – 7.3 h
Life Technologies Ion Torrent Proton PI chip	Semiconductor chip	200 bp	10 Gb	80 million	4 h
Helicos BioSciences Heliscope	Single molecule sequencing	35 bp	35 Gb	1 billion	8 days
Pacific Biosystems PacBio RS	Single molecule real-time sequencing	4000 bp	216 Mb	47,000	0.5 – 2 h

1.3 Metagenomics

Metagenomics is the study of entire communities on the basis of their genetic material from samples obtained directly from the environment, bypassing the requirement for obtaining pure cultures for sequencing (Cardenas and Tiedje, 2008; Hugenholtz and Tyson, 2008; Su *et al.*, 2012). The term "metagenomics" was originally used in 1998, to capture the notion of analysis of a collection of similar but not identical items (Handelsman *et al.*, 1998). Metagenomics can answer questions like "who is there?" to examine community structure; "what can they do?" to determine the genetic potential; "what are they doing" to determine gene expression and function; and "how does it change" to look at what changes occur over time or under different environmental pressures (Cardenas and Tiedje, 2008).

Initial metagenomic analysis involved isolating DNA from an environmental sample, cloning the DNA into a suitable vector, transforming the clones into a host bacterium and screening the resulting transformants, and sequencing the clones that contain phylogenetic information that indicate the probable source of the DNA fragment (Handelsman, 2004). The cultivation bottleneck of traditional microbiology methods has provided a biased view of microbial diversity, but the use of metagenomics allows the full community structure to be seen. As more and more environments have been sampled, it has become evident that the majority of the microbes have yet to be cultured (Cardenas and Tiedje, 2008; Hugenholtz and Tyson, 2008). Sequence analysis of entire microbial communities creates an opportunity to discover a multitude of different bacterial species that are unique to faecal and environmental sources (Dubinsky *et al.*, 2012). DNA-based metagenomics studies frequently fall into one of two categories. Targeted amplicon studies focus on one or a few marker genes and use these markers to reveal the composition and diversity of the microbiota. Other studies use an entire metagenomic approach, or shotgun metagenomics, where entire genomic sequences are generated in a random approach (Kuczynski *et al.*, 2012).

1.3.1 Microbial metagenomics with NGS technology

The rapid and substantial cost reduction in NGS has dramatically accelerated the development of sequence-based metagenomics (Thomas *et al.*, 2012). The first NGS based 16S rRNA study was of deep sea sediments, targeting the V6 hypervariable

region (Sogin *et al.*, 2006). Edwards *et al*. (2006) were also one of the first to use NGS metagenomics for environmental samples, targeting the 16S rRNA gene to determine the microbial communities from groundwater in an iron mine. Since these first pioneering studies, NGS technologies have facilitated mass sequencing of environmental samples from a variety of ecosystems, including freshwater, marine, soil, terrestrial and gut microbiota (Shokralla *et al.*, 2012). Of particular note is the international initiative of the Human Microbiome Project (Turnbaugh *et al.*, 2007), which aims to map human-associated microbial communities, including those of the gut, mouth, skin and vagina (Hugenholtz and Tyson, 2008).

1.3.1.1 Pyrosequencing studies

454 pyrosequencing is the favoured technology for microbial sequencing and metagenomic studies (Cardenas and Tiedje, 2008), with a wide range of communities having been characterised using pyrosequencing methods to date. Bowers *et al.* (2011) characterised 16S rRNA for airborne microbial communities, and determined the influence land usage has on these communities; Cox-Foster *et al.* (2007) surveyed microflora of honey bee colonies to determine the association of microbial community with colony collapse disorder; Huber *et al.* (2007) investigated microbial communities at two hydrothermal vents, targeting the V6 region of the 16S rRNA gene; Li *et al.* (2013) targeted the complete 16S rRNA gene for pyrosequencing of methane-producing microbial communities in a solid-state biogas reactor, and Tripathi *et al.* (2012) sampled tropical soil bacterial communities in Malaysia across a range of land use types and determined community composition through targeting the V1-V3 16S rRNA region. Fierer *et al.* (2007) also studied soil diversities but included four major microbial taxa, bacteria, Archaea, fungi and viruses.

There has been a strong focus on the microflora of humans and animals, with studies including oral microbial communities (Chun *et al.*, 2010), gut microbial communities (Andersson *et al.*, 2008; Degnan *et al.*, 2012; Dethlefsen *et al.*, 2008; Dowd *et al.*, 2008; Flores *et al.*, 2012; Lamendella *et al.*, 2011; Lozupone *et al.*, 2012; Turnbaugh *et al.*, 2009; Wu *et al.*, 2010), skin microbial communities (Fierer *et al.*, 2008; Hulcr *et al.*, 2012; Verhulst *et al.*, 2011) and multiple areas of the human body to determine the spatial and temporal distribution of the human microbiota (Costello *et al.*, 2009; Kuczynski *et al.*, 2010).

1.3.1.2 Illumina studies

Fewer studies have utilised the Illumina platforms, due to the shorter read lengths produced, however, a number of proof-of-concept studies have shown that this platform is still useful. Bartram *et al.* (2011) studied a composite Arctic tundra soil sample through sequencing of the V3 16S rRNA region; Caporaso *et al.* (2012) used mock community samples to demonstrate the sequencing accuracy of the Illumina platform for microbial studies; Lazarevic *et al.* (2009) amplified the V5 16S rRNA region from samples from the oral cavity of three healthy individuals, and in 2012 studied saliva bacterial communities using whole genome shotgun sequencing and regions V1 and V3 of the 16S rRNA gene (Lazarevic *et al.*, 2012); Qin *et al.* (2010) studied the human gut microflora using faecal samples from 124 individuals; and Ye *et al.* (2012) studied the V6 16S rRNA region from activated sludge in wastewater treatment bioreactors.

1.3.1.3 Metagenomic MST studies

There have only been a small number of studies relating to MST to date. Some of these have only looked at faecal source material, including McLellan *et al.* (2010), who looked at eight untreated sewage influent samples from two wastewater plants by targeting the V6 region of the 16S rRNA gene, and Lee *et al.* (2011) who sequenced the V2 region of the 16S rRNA gene for a range of faecal sources, including human, chicken, cow, pig and geese in South Korea. Other studies have focused on water samples. Wilhelm *et al.* (2011) used pyrosequencing of the 16S rRNA gene to determine the community of the Taihu Lake in China, including potential faecal bacteria. However, when these results were compared against qPCR MST methods, there was no indication of faecal contamination. Studies by Unno *et al.* (2010) and Jeong *et al.* (2011) have used 16S rRNA pyrosequencing to investigate the microbial communities of a range of faecal samples and compared them to samples from rivers in South Korea, with results suggesting different contamination sources for different river sites.

1.3.2 Barcoding strategies

In many metagenomic projects, the number of sequencing reads that are generated through a NGS sequencing run vastly exceeds the number of reads required for the samples being analysed. To maximise the high throughput capabilities of these NGS

platforms, most metagenomics projects benefit from pooling multiple PCR amplicon samples together for sequencing, often termed ‘multiplex sequencing’ or barcoding. In order to do this, the results from each sample must be able to be identified. This is achieved through the addition of sample-specific barcodes to the amplicons during the library preparation steps. The adoption of nucleotide barcodes in amplification primers allows samples from different origins to be mixed in one run, and their data easily separated out according to their barcode after sequencing (Cardenas and Tiedje, 2008; Meyer *et al.*, 2007; Parameswaran *et al.*, 2007). This decreases the cost per sample as more samples can be pooled in a single sequence run, rather than sequencing fewer samples to greater depth. For example, using a 454 GS Junior sequencer, 100,000 reads are usually generated; by using 20 barcodes 5,000 reads per sample can be obtained, a value which is at least one order of magnitude higher than those from traditional clone libraries (Cardenas and Tiedje, 2008). Multiplexing in amplicon sequencing can be performed either by ligating barcodes and sequencing adaptors to amplicons created with conventional PCR primers (Meyer *et al.*, 2008), or by using fusion primers with the barcode incorporated, thereby eliminating the ligation step (Binladen *et al.*, 2007; Huse *et al.*, 2010).

Barcode sequences have a couple of requirements for design. They need to be relatively short to save most of the limited sequencing read length for the sample sequence, but also long enough to allow the required numbers of samples to be sequenced concurrently and to be substantially different from each other to prevent cross mutation between sample tags (Bystrykh, 2012). Early barcoding design utilised extremely short sequences, only 2 or 4 nt in length (Binladen *et al.*, 2007; Hoffmann *et al.*, 2007), however, it has been suggested that these are too short to allow for massively parallel runs involving large numbers of sample libraries, and have a steeper trade-off between number of possible barcodes and the minimum number of nucleotide differences between individual barcodes (Parameswaran *et al.*, 2007). Most NGS sequencing platforms have designed their own multiplex tags, such as the 10 nt Roche Multiplex Identifiers (MIDs) and the 6 nt Illumina identifier tags. These have been designed to make the most of the flow cycle set up of their respective platform. Other suggested barcodes have been designed using coding theory (Bystrykh, 2012; Hamady *et al.*, 2008; Krishnan *et al.*, 2011), which provides an inbuilt component of error-correcting.

1.4 Objectives of this study

The main objective of this study was to investigate the application of next generation sequencing platforms in the field of microbial source tracking within New Zealand. The most commonly used method for microbial source tracking is utilising PCR to amplify source-specific markers, usually from the *Bacteroidetes* phyla, using the 16S rRNA gene as the target. This gene has been shown to have enough genetic diversity to classify bacteria to the species level, and does not require the whole gene to be targeted to infer taxonomy. Next generation sequencing methods also frequently target the 16S rRNA gene, but rather than targeting individual species, by using universal bacterial primers, total bacterial 16S rRNA gene sequences can be amplified and sequenced.

The strategy in this project was to sample faecal material from a range of animals and birds known to contribute to the faecal contamination of New Zealand waterways, and generate sequence databases comprised of individual libraries of partial 16S rRNA gene sequences of bacteria from each faecal source. Bacterial sequences can be compared to readily accessible 16S rRNA gene databases, resulting in taxonomic identification of the sequences. Statistical analysis can be used to determine the microbial diversity of the different faecal source types, looking at both intra-sample diversity (alpha diversity) and inter-sample diversity (beta diversity).

In the same manner, water samples from waterways that are thought to be polluted with faecal contamination can be studied through sequencing of the 16S rRNA gene. Sequences generated by next generation sequencing can be compared against the database of known bacterial species found in different faecal sources. Statistical analyses can be applied to determine the likelihood of each water sample being contaminated by the known faecal sources.

Chapter Two
Optimisation of protocols and analysis

2.1 Abstract

Microbial source tracking (MST) is used to determine the source of faecal contamination in waterways, which is important for identifying bacterial pathogens associated with human and animal diseases. MST methods assist with developing effective management strategies for controlling and eliminating the source of the pollution. However, traditional methods for determining faecal contamination, which measure faecal indicator bacteria, may not identify all faecal sources. Recent advances in DNA sequencing technologies enable rapid sequencing of a large number of nucleic acid sequences from multiple environmental samples at once, through the utilisation of sample-specific barcodes. As a proof-of-concept MST study, for use within New Zealand, a protocol was optimised for preparing 16S rRNA gene amplicons from faecal and water samples for next generation sequencing. A range of currently available analysis tools were trialled. A total of 10,409 sequences were generated, of which 8,358 were taxonomically classified using the Ribosomal Database Project Classifier, and 7,777 were taxonomically classified through the QIIME pipeline. The majority of bacteria in the faeces of sheep, cows and ducks were *Bacteroidetes* and *Firmicutes*; human sewage samples were dominated by *Proteobacteria*, and swan faecal samples with *Fusobacteria*, whereas all three water samples were dominated by *Proteobacteria*. Community diversity analyses were conducted using the QIIME platform, which generated rarefaction curves, Principal Coordinate Analysis plots and bootstrapped phylogenetic trees. These support clustering of the sheep and cow samples together, with the human samples also forming their own specific cluster. The water samples were not found to cluster closely with any of the faecal source samples. The methods presented here provide a suitable application of next generation sequencing methods to microbial source tracking in New Zealand, and will contribute to further development in this area.

2.2 Introduction

There are a large number of water bodies throughout the world which are considered to be impaired on the basis of their microbiological qualities, usually due to faecal contamination. Traditional faecal assessment methods have studied faecal indicator bacteria, such as *E. coli, enterococci* and culturable coliforms (Hagedorn and Liang, 2011; Tallon *et al.*, 2005), however, the use of these organisms does not reveal the source of the contamination, as they are found in the faeces of a variety of warm- and cold-blooded animals (Field and Samadpour, 2007). Identifying the faecal source is required to ensure elimination of the pollution and to minimise the impact on human disease. A range of microbial source tracking (MST) methods have been developed as a means of identifying the source of contamination, based on phenotypic and genotypic methods. Of these, targeting the 16S rRNA gene as a marker through the use of PCR has received the most attention (Cardenas and Tiedje, 2008; Clarridge III, 2004). A large number of specific primers have been designed that target source-specific sequences, including humans, ruminants, pigs, dogs, gulls and geese (Roslev and Bukh, 2011). There are also a number of "universal" primers, which target the conserved regions of the 16S rRNA gene, allowing all prokaryotic 16S rRNA genes in a given sample to be amplified (Baker *et al.*, 2003). These primer pairs tend to target multiple hypervariable regions of the 16S rRNA gene, which provide enough species-specific variation for taxonomic identification between the organisms at the genus level, across all phyla of bacteria (Chakravorty *et al.*, 2007).

Recent advances in DNA sequencing technologies have resulted in the ability to rapidly sequence large amounts of DNA at a reasonable cost through the use of next generation sequencing (NGS) platforms. This has led to a large number of environmental metagenomic studies, where entire bacterial communities have been analysed in a variety of different environments (Shokralla *et al.*, 2012). Most of these studies have utilised the 454 pyrosequencing platforms, as this technology offers the longest read lengths of all the NGS platforms available, with average read lengths approximately 500 nt and well characterised error rates (Valverde and Mellado, 2013). The advantage of this for environmental DNA sequencing is that PCR amplicons targeting a region of the 16S rRNA gene can be sequenced from a single direction, which provides enough sequence data for taxonomic classification (Claesson *et al.*, 2009; Mizrahi-Man *et al.*, 2013). The ability to barcode samples and pool multiple samples together in a single

sequencing run allows the high throughput capabilities of these NGS platforms to be maximised to their full potential by being able to sequence potentially hundreds of samples simultaneously. By including sample-specific barcode sequences during the sample preparation steps, sequences generated for each sample can easily be separated out based on their barcode after sequencing (Cardenas and Tiedje, 2008; Parameswaran *et al.*, 2007).

Metagenomic studies utilising the 16S rRNA gene have either sequenced the full gene through shotgun sequencing, or sequenced targeted amplicons, where only a portion of the gene is sequenced (Kuczynski *et al.*, 2012). Full length sequencing offers higher degrees of taxonomic resolution, while targeted amplicon sequencing allows a greater sampling depth, facilitating the investigation of less-abundant taxa which may otherwise be missed. Almost all metagenomics studies utilising NGS platforms have used 16S rRNA amplicons, targeting a range of hypervariable regions, including V1-V2 (Costello *et al.*, 2009; Fierer *et al.*, 2008; Lee *et al.*, 2011), V1-V3 (Carroll *et al.*, 2012; Chun *et al.*, 2010; Unno *et al.*, 2010), V3 (Dethlefsen *et al.*, 2008), V3-V4 (Flores *et al.*, 2012), V3-V5 (Wu *et al.*, 2010), V4 (Claesson *et al.*, 2009; Hulcr *et al.*, 2012), V4-V5 (Zhou *et al.*, 2011), V4-V6, (Dowd *et al.*, 2008), V5-V6 (Andersson *et al.*, 2008; De Filippo *et al.*, 2010), V6 (Claesson *et al.*, 2009; Dethlefsen *et al.*, 2008; McLellan *et al.*, 2010), and V6-V9 (Degnan *et al.*, 2012; Wu *et al.*, 2010) (refer to Figure 1.1). To date, no single region has received universal acceptance (Schloss *et al.*, 2011). However, the V1-V3 region has been shown to result in a deeper richness (Handl *et al.*, 2011), provide a higher degree of classification accuracy (Kim *et al.*, 2011; Wang *et al.*, 2007) and a lower degree of classification bias towards specific taxonomic groups (Vilo and Dong, 2012), compared to other regions. Because sequencing of the rRNA gene has become the method of choice for investigating microbial diversity, there are vast amounts of rRNA gene sequence data available in public databases (Quast *et al.*, 2013). Four 16S rRNA-specific databases are currently available, the Ribosomal Database Project (RDP) (Cole *et al.*, 2009), Greengenes (DeSantis *et al.*, 2006), SILVA (Quast *et al.*, 2013) and EzTaxon-e (Kim *et al.*, 2012), each with their own advantages and disadvantages.

The ability to analyse the vast quantities of sequence data generated by NGS platforms has led to the need for easily accessible, well documented and well tested tools, particularly in the form of a pipeline which provide complete analysis solutions (Gonzalez and Knight, 2012). A number of individual programme packages are

available as open source software, generally designed by specific laboratories to fit their requirements, including alignment (Caporaso *et al.*, 2009; Nawrocki *et al.*, 2009; Tamura *et al.*, 2011), clustering and phylogenetic analyses (Edgar, 2010; Li and Godzik, 2006; Lozupone *et al.*, 2006; Price *et al.*, 2010). Other programmes have been designed to incorporate multiple components of analyses, such as ARB (Ludwig *et al.*, 2004), MEGAN (Huson *et al.*, 2011), Mothur (Schloss *et al.*, 2009) and QIIME (Caporaso *et al.*, 2010). RDP also includes a pyrosequencing pipeline (Cole *et al.*, 2009), which provides many of these analysis options as well. As with the 16S rRNA databases, no single analysis platform has emerged as the best; however, the ability to follow data from raw to complete analysis using a single programme, such as QIIME and Mothur, currently provide the simplest way to analyse large quantities of sequencing data without a large background knowledge of statistical programmes required.

Only a limited number of studies have used NGS techniques for MST. Unno *et al.* (2010) used 454 pyrosequencing to target the V1-V3 region of the 16S rRNA gene in a range of faecal and animal sources. For their data analyses they used a range of programmes, including CD-HIT, Mothur, Ez-Taxon-e and MEGA, which ultimately provided a density ratio for each faecal contamination source in their water samples. Jeong *et al.* (2011) also targeted the V1-V3 region for a range of faecal and water samples, using the RDP Classifier for assigning taxonomy, followed by alignment with the RDP pyrosequencing pipeline and they used the Mothur programme to estimate bacterial diversity. Lee *et al.* (2011) characterised the microbial compositions of human and animal faeces as sources of faecal contamination for MST studies, by sequencing the V1-V2 region of the 16S rRNA gene. They utilised the QIIME platform, including alignment with PyNAST, taxonomic classifications with the RDP Classifier, and Principal Coordinate Analysis (PCoA) using the Unique Fraction metric (UniFrac).

As a proof-of-concept study, we generated a small NGS dataset, using a range of faecal and water samples from New Zealand, to assess the use of high-throughput sequencing in New Zealand for MST applications. A DNA extraction and NGS sample preparation protocol was optimised and evaluated through a small number of the data analysis programs freely available. The V1-V3 hypervariable region of the 16S rRNA gene was selected as the target because of the large number of successful universal primers available for this region. We selected the Roche 454 pyrosequencing platform as the longer read lengths generated make it ideal for amplicon analysis. This library

preparation and data analysis protocol provides a proof-of-concept study for applying NGS technologies and bacterial diversity analysis to MST techniques available in New Zealand.

2.3 Methods and materials

2.3.1 Sample preparation

2.3.1.1 Faecal sample collection and DNA extraction

Fresh faecal samples from a known species source, often identified immediately after observed defecation, were collected in sterile containers, avoiding contact with grass and soil as much as possible. Faecal samples were transferred to the laboratory in a cooled chilly bin as soon as possible after collection. Where overnight storage was required prior to DNA extraction procedures, samples were kept at 4°C. Details of faecal samples collected are provided in Table 2.1.

Each sample was processed individually, with 1 g faecal material added to 4.0 ml GITC buffer (5 M Guanidine isothiocyanate; 100 mM EDTA; 0.5% Sarcosyl), and the resulting slurry stored at either 4^{o}C or -80^{o}C for a minimum of 2 h and up to maximum of 24 h. Total genomic DNA extraction was carried out using either DNA-EZ RW02 Kit (GeneRite, USA) or ZR Fecal DNA Miniprep Kit (Zymo Research Corp, USA). All buffers used during the extractions were provided by the respective extraction kit.

2.3.1.1.1 GeneRite extraction protocol

400 µl Elution Buffer at 60°C was added to a provided bead tube, with 300 µl of faecal slurry. Samples were homogenised using a Mixmate beater (Eppendorf, Germany) for 10 min at 2,000 rpm, and incubated for 1 h at room temperature. Following this, the bead tubes were centrifuged for 1 min at 13,200 rpm, the supernatant transferred to a new 1.5 ml eppendorf tube, and centrifuged for a further 1 min at 13,200 rpm. 500 µl of supernatant was transferred to a new eppendorf tube, with 1,000 µl Binding Buffer, and vortexed to mix. 750 µl of this lysate was transferred into a DNAsure column with collection tube, and centrifuged for 1 min at 13,200 rpm. The collection tube contents were emptied and this step was repeated after the addition of the remaining lysate to the DNAsure column. 500 µl EZ-Wash Buffer was added to the column, and centrifuged at 13,200 rpm for 1 min, followed by a second wash step. The column was centrifuged for

a further 2 min at 13,200 rpm to ensure no trace of ethanol from the Wash Buffer was left. The column was transferred to a new 1.5 ml eppendorf tube, and 100 µl of Elution Buffer at 60^{o}C was added directly to the column membrane, and incubated at room temperature for 5 min. DNA was eluted into the eppendorf tube by centrifugation for 1 min at 13,200 rpm. Extracted DNA was stored at 4^{o}C.

Table 2.1: Faecal library samples used in the GS454-01 sequencing study. * indicates samples freshly collected and extracted for this study. Other samples were archived extracted DNA samples stored at 4^{o}C. Composite samples containing DNA extracted from five individual samples were prepared after DNA extraction. Human sewage samples were not composited, due to already containing a mixed human faecal source. qPCR results are the threshold cycle (Cp) values for a general-source qPCR assay.

ESR Sample ID	Previous qPCR results	Species	Location sample taken from	Study Sample ID	Barcode tag
CMB05176 CMB05177 CMB05178 CMB05179 CMB05180		Sheep (*Ovis aries)*	Lyttelton	NGS001	B4.1
CMB05188 CMB05189 CMB05190 CMB05191 CMB05192	15.09 14.35 14.86	Sheep (*Ovis aries)*	Dunsandel	NGS002	B4.2
CMB120037* CMB120038* CMB120039* CMB120040* CMB120041*	13.83 14.50 14.60 15.16 14.68	Sheep (*Ovis aries)*	Christchurch	NGS003	B4.3 B4.9
CMB120325* CMB120326* CMB120328* CMB120331* CMB120333*	15.03 14.74 14.61 14.76 15.75	Sheep (*Ovis aries)*	Winchmore	NGS004	B4.4
Cawthron 7 Cawthron 8 Cawthron 9 Cawthron 10 Cawthron 11	20.75 20.67 20.37 21.60 19.62	Cow (*Bos primigenius)*	South Island	NGS005	B4.5 B4.11
CMB06648 CMB06649 CMB06650 CMB06651 CMB06680	22.4 22.0 21.9 21.0 16.7	Cow (*Bos primigenius)*	Cust	NGS006	B4.6
CMB06684 CMB06685 CMB06686 CMB06687 CMB06688	17.86 16.2	Cow (*Bos primigenius)*	Lincoln	NGS007	B4.7

Table 2.1 continued

ESR Sample ID	Previous qPCR results	Species	Location sample taken from	Study Sample ID	Barcode tag
MB1004001 MB1004002 MB1004003 MB1004004 MB1004005		Cow (*Bos primigenius)*	Hanmer Springs	NGS008	B4.8
CMB05221 CMB05222 CMB05223 CMB05224 CMB05225	40 40 40 36.48	Seagull (*Larus spp.)*	Sumner beach, Christchurch	NGS009	
CMB05230 CMB05231 CMB05232 CMB05233 CMB05234	14.6	Duck (Anatidae)	Hagley Park, Christchurch	NGS010	B4.10
CMB091261 CMB091262 CMB091263 CMB091264 CMB091268	 37.37	Canada Geese (*Branta Canadensis)*	Bromley, Christchurch	NGS011	
CMB09197 CMB09198 CMB09199 CMB09200 CMB092001	16.07 25.58	Swan (*Cygnus)*	Bromley, Christchurch	NGS012	B4.12
Cawthron 119	16.33	Human sewage	Northland	NGS013	B4.13
CMB05123	21.9	Human sewage	Bromley	NGS014	B4.14
CMB06668	22.2	Human sewage	Bromley	NGS015	B4.15

2.3.1.1.2 Zymo extraction protocol

300 µl faecal slurry was added to a provided ZR Bashing Bead Lysis Tube, with 750 µl Lysis Solution. Samples were homogenised using a Mixmate beater (Eppendorf, Germany) for 10 min at 2,000 rpm, followed by centrifugation for 1 min at 13,200 rpm. 400 µl of supernatant was transferred to a Zymo-Spin IV filter in a collection tube, and centrifuged for 1 min at 7,000 rpm. 1,200 µl Fecal DNA Binding Buffer was added to the filtrate in the collection tube. 800 µl of this mixture was transferred to a Zymo-Spin IIC column and centrifuged for 1 min at 13,200 rpm. The flowthrough was discarded and the step repeated with the remaining 800 µl of filtrate/Binding Buffer mixture. 200 µl of DNA Pre-Wash Buffer was added to the Zymo-Spin IIC column and centrifuged for 1 min at 13,200 rpm. 500 µl of Fecal DNA Wash Buffer was added to the column and centrifuged for 1 min at 13,200 rpm. The column was transferred to a new 1.5 ml eppendorf tube and 100 µl DNA Elution Buffer was added directly to the column matrix, followed by centrifugation for 30 s at 13,200 rpm. The eluted DNA was

transferred to a prepared Zymo-Spin IV-HRC Spin filter in a new 1.5 ml eppendorf tube and centrifuged for 1 min at 8,000 rpm to elute the final DNA extraction. Extracted DNA was stored at 4^{o}C.

2.3.1.1.3 Quantification of genomic DNA

Quantitation and purity of DNA was determined by absorption spectroscopy, using a Nanodrop spectrophotometer ND-1000 (Thermo Scientific Inc.). A 1.5 µl blank of the elution buffer used for the extraction of the DNA sample was initially used to zero the spectrophotometer before determining the absorbance at 260 nm (A_{260}). The purity of the nucleic acid sample was estimated from the A_{260}/A_{280} ratio. A ratio of 1.8 to 2.0 indicated a highly purified preparation of DNA that was suitable for further sequencing preparation.

2.3.1.2 Environmental water sample preparation

All water samples used in this study were samples that had been provided to ESR for MST work. Information from the commercial analysis previously undertaken by ESR was used to select suitable samples to ensure a range of contamination levels from different sources was analysed by this study (Table 2.2).

Table 2.2: Water samples used in the GS454-01 sequencing study.

ESR Sample ID	Location sample taken from	Previous ESR contamination analysis outcome	Study Sample ID	Barcode tag
CMB120274	Auckland	Human	NGS016	B4.16
CMB120322	Northland	Ruminant	NGS017	B4.17
CMB120397	Southland	Ruminant	NGS018	B4.18

2.3.2 Sequencing library preparation

Libraries were prepared from both archived DNA samples and from freshly collected and extracted samples (Table 2.1). Archived DNA samples were selected using information on date collected, location, and previously obtained threshold cycle (Cp) quantitative PCR (qPCR) results, to ensure enough individual samples collected from the same location at a similar time were available to be pooled.

2.3.2.1 Pooling of DNA extraction samples

For most species, DNA extraction samples were pooled to obtain a composite sample (Table 2.1). 5 µl from each of five individual samples from the same species, from the same collection location and similar collection dates, were pooled together. If previous qPCR information was available, such as GenBac3 Cp values, these were taken into consideration in an attempt to keep individual samples in equivalence. Where a Cp value differed from others by approximately 3, this indicated the sample to have a 1 log concentration difference, and was diluted accordingly prior to pooling. Pooled DNA samples were diluted in molecular biology grade water (UltraPure Distilled Water, Invitrogen) for further analysis, either to a 1:10 or a 1:100 dilution.

For human raw sewage samples, which are already a composite of multiple individuals, no pooling was required. Likewise, environmental water samples were not pooled, but left as individual samples.

2.3.2.2 Amplicon preparation

2.3.2.2.1 Oligonucleotide primer design

The selection of the primers used in this study was based on literature also using 16S rRNA gene targets. The variable regions 1 to 3 (V1-V3) were selected as the target region, as previous studies have shown these regions to be suitable for distinguishing between most bacterial species (Chakravorty *et al.*, 2007). The final primer selection was from Fierer *et al.* (2007), using a universal eubacterial primer set, Bac8F (5'-AGAGTTTGATCCTGGCTCAG-3') and Univ529R (5'-ACCGCGGCKGCTGGC-3'), resulting in an amplicon approximately 520 nt in length (Figure 2.1).

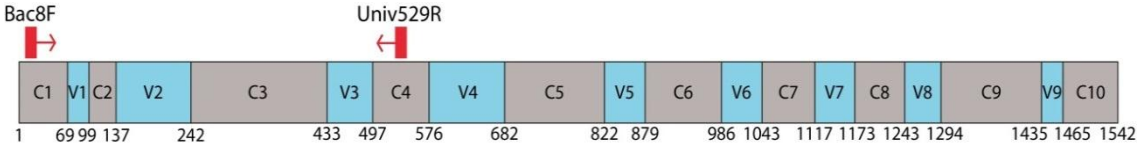


Figure 2.1: Binding sites of 16S rRNA Bac8F and Univ529R primers. The conserved regions of the 16S rRNA gene (C1-C10) are shown in grey and the hypervariable regions (V1-V9) in blue. Nucleotide positions are based on *E. coli* nomenclature. The binding sites of the V1-V3 region primers used to produce the amplicon are shown in red.

Barcoding was utilised to enable multiple samples to be sequenced together in a manner that would allow easy identification of where the corresponding sequences originated from. Forward and reverse primers each contained a four nt barcode sequence at the 5' end of the corresponding primer (Roossinck *et al.*, 2010). The addition of the barcodes to the primers results in an amplicon approximately 530 nt in length. All primers were purchased from Invitrogen, and are listed in Table 2.3.

Table 2.3: PCR primers used for samples in GS454-01 sequencing study. Barcode sequences are highlighted in bold.

Primer	Nucleotide sequence	T_m (°C)
Bac8F-B4.1	**AGAG**AGAGTTTGATCCTGGCTCAG	52
Univ529R-B4.1	**AGAG**ACCGCGGCKGCTGGC	56
Bac8F-B4.2	**ACTC**AGAGTTTGATCCTGGCTCAG	52
Univ529R-B4.2	**ACTC**ACCGCGGCKGCTGGC	56
Bac8F-B4.3	**AGTG**AGAGTTTGATCCTGGCTCAG	52
Univ529R-B4.3	**AGTG**ACCGCGGCKGCTGGC	56
Bac8F-B4.4	**ATAG**AGAGTTTGATCCTGGCTCAG	51
Univ529R-B4.4	**ATAG**ACCGCGGCKGCTGGC	54
Bac8F-B4.5	**ACAC**AGAGTTTGATCCTGGCTCAG	52
Univ529-B4.5	**ACAC**ACCGCGGCKGCTGGC	56
Bac8F-B4.6	**CACA**AGAGTTTGATCCTGGCTCAG	52
Univ529-B4.6	**CACA**ACCGCGGCKGCTGGC	56
Bac8F-B4.7	**CTCT**AGAGTTTGATCCTGGCTCAG	52
Univ529R-B4.7	**CTCT**ACCGCGGCKGCTGGC	56
Bac8F-B4.8	**CAGA**AGAGTTTGATCCTGGCTCAG	52
Univ529R-B4.8	**CAGA**ACCGCGGCKGCTGGC	56
Bac8F-B4.9	**CTGT**AGAGTTTGATCCTGGCTCAG	52
Univ529-B4.9	**CTGT**ACCGCGGCKGCTGGC	56
Bac8F-B4.10	**ATGC**AGAGTTTGATCCTGGCTCAG	52
Univ529R-B4.10	**ATGC**ACCGCGGCKGCTGGC	56
Bac8F-B4.11	**GAGA**AGAGTTTGATCCTGGCTCAG	52
Univ529R-B4.11	**GAGA**ACCGCGGCKGCTGGC	56
Bac8F-B4.12	**GTGT**AGAGTTTGATCCTGGCTCAG	52
Univ529R-B4.12	**GTGT**ACCGCGGCKGCTGGC	56
Bac8F-B4.13	**GACA**AGAGTTTGATCCTGGCTCAG	52
Univ529R-B4.13	**GACA**ACCGCGGCKGCTGGC	56
Bac8F-B4.14	**GTCT**AGAGTTTGATCCTGGCTCAG	52
Univ259R-B4.14	**GTCT**ACCGCGGCKGCTGGC	56
Bac8F-B4.15	**GATC**AGAGTTTGATCCTGGCTCAG	52
Univ529R-B4.15	**GATC**ACCGCGGCKGCTGGC	56
Bac8F-B4.16	**TCTC**AGAGTTTGATCCTGGCTCAG	52
Univ529R-B4.16	**TCTC**ACCGCGGCKGCTGGC	56
Bac8F-B4.17	**TGTG**AGAGTTTGATCCTGGCTCAG	52
Univ529R-B4.17	**TGTG**ACCGCGGCKGCTGGC	56
Bac8F-B4.18	**TCTG**AGAGTTTGATCCTGGCTCAG	52
Univ529R-B4.18	**TCTG**ACCGCGGCKGCTGGC	56

2.3.2.2.2 PCR amplification of DNA targets

Prior to PCR set up, stock primers (100 pmol/µl) were diluted 1:10 in molecular biology grade water to give a final concentration of 10 pmol/µl. Amplification of specific targets was performed in a 50 µl reaction mixture, in 0.2 ml thin-walled PCR tubes (Scientific Specialties Inc., USA) in an automatic thermal cycler (GeneAmp PCR System 9700, Applied Biosystems). Platinum Taq High Fidelity Polymerase enzymes were purchased from Invitrogen (USA) and dNTPs were obtained from Life Technologies (USA).

PCR reaction mixes were made up as master mixes, containing PCR buffer, Mg^{2+}, dNTPs, HiFi DNA polymerase enzyme and water, as listed in Table 2.4. The primers were individually barcoded, therefore added to each PCR reaction separately.

Table 2.4: PCR reaction mix for samples in the GS454-01 sequencing study.

Reagent	Concentration per reaction tube	Volume per reaction tube (µl)
10x Buffer	1x	5
Mg^{2+} (50 mM $MgSO^4$)	2 mM	2
dNTPs (25 mM each)	0.2 mM each	0.4
HiFi Polymerase (5 units/µl)	1 unit	0.2
Primers (10 pmol/µl)	0.2 µM each	1 (of each)
DNA		2
dH_2O		38.4
Total		50

PCR amplification was initiated with a denaturation step at 96°C for 4 min, followed by a three stage programme of 30 repeated cycles. Each amplification cycle consisted of a denaturation step (95°C for 30 s), an annealing step (55°C for 30 s) and an extension step (68°C for 30 s). A second extension step of 68°C for 10 min followed these 30 cycles. A final step of 20°C was included to keep reactions at room temperature until processing. PCR products were stored at 4°C.

The resulting PCR products were run on a 2% agarose gel subjected to electrophoresis at 110 V for 1 h in TBE buffer containing ethidium bromide (EtBr). Gels were visualised under UV light to determine the presence and size of PCR amplicons.

Multiple amplification rounds were used for samples which did not amplify well, in particular the seagull sample, NGS009 and the Canadian geese sample, NGS011. Various attempts were made to optimise the conditions for these birds, including

altering the magnesium levels and changing the PCR protocol to a two-temperature cycle, where the denaturation steps were left as above, with the annealing and extension temperatures combined to 68°C for 1 min. The number of cycles was kept at 30, with a final extension step at 68°C for 10 min. This protocol worked well for some samples, in particular sheep NGS003 and cow NGS005. These two samples were included in the sequencing run to determine the effects of the two-temperature protocol on amplified sequences (NGS003.B4.3 and NGS005.B4.5), but the protocol was not adopted for all samples.

We were unable to optimise a protocol that would amplify NGS009 and NGS011 well enough to use for sequencing, so these samples were not included in the final sequencing sample.

2.3.2.3 Purification of PCR amplicons

2.3.2.3.1 AMPure XP purification

All positive PCR reactions, as determined by gel electrophoresis, were purified using Agencourt AMPure XP magnetic beads (Beckman Coulter, USA), following the manufacturer's directions, with some modifications as suggested in the Roche 454 Amplicon Library Preparation Method Manual for GS FLX Titanium series. Briefly, 35 µl of PCR amplicon was transferred into a new round bottom PCR plate (Scientific Specialties Inc., USA) with 15 µl molecular biology grade water and 72 µl AMPure XP beads. The solution was gently mixed by pipetting and incubated at room temperature for 5 – 10 min. The PCR plate was placed into an Agencourt SPRIPlate Super Magnet Plate (Beckman Coulter, USA) for 2 – 5 min to separate the magnetic beads from the solution. The supernatant was removed and discarded while the PCR plate was on the magnetic plate. 200 µl of freshly prepared 70% ethanol was added to each amplicon sample, and incubated at room temperature for 1 min, then removed and discarded. This wash was repeated for a total of two washes. The PCR plate was left to dry at room temperature for up to 5 min to ensure all traces of ethanol were removed. The PCR plate was removed from the magnetic plate, 40 µl of TE buffer added to each amplicon sample, mixed gently by pipetting and incubated at room temperature for 2 min. The PCR plate was placed back onto the magnetic plate for a further 1 min to separate the beads from the solution, and the solution removed and transferred to a new 0.5 ml eppendorf tube. Purified amplicons were stored at -20^{o}C until required for sequencing.

2.3.2.3.2 Quantification of PCR amplicons

PCR amplicons were quantified using the Qubit dsDNA HS Assay kit (Invitrogen, USA) with the Qubit 1.0 Fluorometer (Invitrogen, USA), following the manufacturer's directions. Briefly, 10 µl of each of the two Qubit standards supplied was added to 190 µl of Qubit working solution and mixed by vortexing. 2 µl of each PCR amplicon was added to 198 µl Qubit working solution and mixed by vortexing. All tubes were incubated at room temperature for 2 min prior to quantification using the dsDNA High Sensitivity assay type in the Qubit fluorometer. The two standards were used initially to calibrate the fluorometer.

2.3.2.3.3 Pooling of amplicons

Amplicons were diluted in molecular biology grade water to a concentration of 5 ng/µl, and 5 µl of each diluted amplicon was pooled together to give a total concentration of 500 ng per sequencing run. Pooled amplicons were stored at -20^{o}C until required for shipping to the sequencing provider. 5 µl of each 5 ng/µl amplicon product was run on an agarose gel as per the conditions above, and the image provided to the sequencing provider.

2.3.3 Next generation sequencing

Sequencing service of the pooled PCR products, including two aquifer samples not analysed as part of this thesis, was provided by New Zealand Genomics Ltd (NZGL), and performed by Auckland University's Centre for Genomics, Proteomics and Metabilomics, on a Roche 454 GS Junior platform. A final library preparation step was performed by Auckland University, where the Roche 454 sequencing adaptors were ligated on to the DNA sequences in the sample using the Roche. This was followed by an emPCR amplification step and sequencing. A second sample, containing the same 20 pooled amplicons, with approximately 500 ng, was later sent for re-analysis, due to problems with the ligation step at Auckland University.

2.3.4 Data analysis

A variety of software available for processing NGS data were used to process and analyse the raw data.

2.3.4.1 Geneious

Geneious R6 (Biomatters, New Zealand) is a Windows-based software package available at http://www.geneious.com/. It incorporates a number of next generation sequencing analysis tools, including sorting sequences by barcode sequence, assembly and mapping abilities, and sequence alignment.

2.3.4.1.1 Initial processing of data

The raw data file was imported into Geneious, and regions identified by the 454 sequencing software as poor quality were removed; this also removed the GACT "key" region required for 454 sequencing. Sequences were filtered by searching for a perfect match to the 454 Rapid Library MID adaptor added by Auckland University, the Roche 454 Rapid Library MID 1 adaptor (ACACGACGACT). This created two new sequence libraries, one with all sequences containing a perfect match to the sequencing adaptor in the first 11 bases, and another with all the sequences that did not contain a perfect adaptor match, which were not used for further analysis. Sequences with a perfect match to the 454 sequencing adaptor were then sorted into different lists based on the sample barcode, comprised of the first four nucleotides of each sequence. The 20 barcodes were loaded into Geneious and the sequences compared against them, creating 21 new sequence libraries, one library for each barcode and a final library containing all sequences with no perfect matching barcode sequence. The barcode sequences were removed so that this sequence would not interfere with alignment and classification steps later in the analysis. The final step in the initial processing was to remove short sequences. A length of 200 bp was selected, based on other literature and information on the Ribosomal Database Project (RDP) website (http://rdp.cme.msu.edu/). Figure 2.2 shows the basic workflow for the initial processing of raw data.

2.3.4.1.2 Primer filtering and alignment

Geneious was used to search for the forward and reverse 16S rRNA amplification primer sequences (AGAGTTTGATCCTGGCTCAG and ACCGCGGCKGCTGGC), by adding each primer sequence as a motif annotation. Sequences were then searched against these motifs with no mismatches allowed, and added as an annotation at the location of the correct primer match. A number of the sequences had been sequenced from the 3' end, so these sequences were reverse complemented in Geneious to ensure all sequences were in the same orientation. Sequences that had no correct match for either the forward or the reverse primer, two of the same primer sequence, or sequences

with primers found within the middle section of sequence were removed from the analysis. The selected sequences for each sample were individually aligned within Geneious using a MUSCLE alignment with a maximum of 8 iterations. This was used as a visual confirmation that all sequences were in the same orientation. Each filtered sample list was exported as a new FASTA file for further analysis.

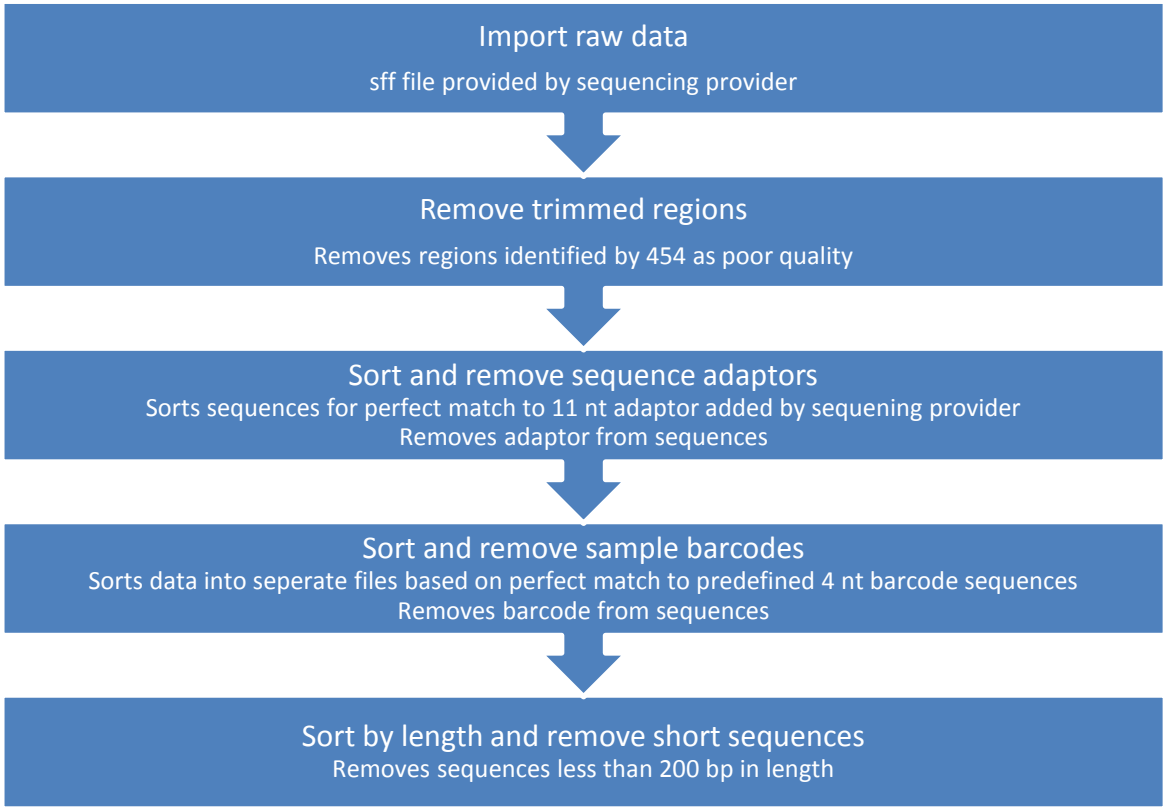


Figure 2.2: Geneious workflow for initial processing of data.

2.3.4.2 Ribosomal Database Project

The Ribosomal Database Project (Cole *et al.*, 2009) is an online database providing ribosome related data services. RDP 10.31 was used for this study, which consists of 2,639,157 aligned and annotated 16S rRNA sequences.

2.3.4.2.1 Pyrosequencing pipeline initial process

Initially data were uploaded via the RDP's pyrosequencing pipeline, which provides a number of tools to take raw pyrosequencing data and perform a range of analyses and convert the data into formats suitable for other statistical packages. In order for the initial process pipeline to sort the raw data into samples, a tab delimited file containing information on the barcodes used must be supplied by the user, as well as the primer

sequences used. This step was unsuccessful with the GS454-01A data, so was not used again.

2.3.4.2.2 RDP Classifier

The RDP Classifier (Wang *et al.*, 2007) assigns 16S rRNA sequences to the new phylogentically consistent higher-order bacterial and fungal taxonomy, based on the RDP naïve Bayesian rRNA Classifier, using the RDP 16S rRNA training set 9 (Cole *et al.*, 2009). Because the sequences were all partial sequences, a bootstrap cutoff confidence threshold of 60% was used for classifying, as it has previously been shown that a bootstrap cutoff of 50% or greater is sufficient to accurately classify sequences at the genus level for partial sequences of length shorter than 250 bp (Claesson *et al.*, 2009).

The FASTA file from each library generated in Geneious was uploaded into the RDP Classifier. Two files were produced by the analysis, a hierarchy file giving the total number of sequences assigned to each classification and an 'allrank' file, which provided the results for all classification levels applied to each sequence. Each file was exported as a comma separated value file and analysed in Microsoft Excel.

2.3.4.3 QIIME: Quantitative Insights Into Microbial Ecology

QIIME (Caporaso *et al.*, 2010) is a Linux-based open source software package designed for comparison and analysis of microbial communities data obtained from NGS amplicon sequencing. QIIME provides a pipeline that takes raw sequencing data through filtering of data and demultiplexing, initial analyses, such as picking operational taxonomic units (OTUs), taxonomic assignment against established databases, such as the RDP classifier, and construction of phylogenetic trees. It also provides statistical analyses and visualisations of this data, such as rarefaction curves and diversity plots. QIIME makes use of other open source tools as part of many of its pipeline processes, including Uclust (Edgar, 2010), RDP classifier (Wang *et al.*, 2007), PyNAST (Caporaso *et al.*, 2009) and FastTree2 (Price *et al.*, 2010).

A full list of scripts used is provided in Appendix I.

2.3.4.3.1 Setting up QIIME

QIIME 1.6.0 was set up on an 8-core Windows 2008 R2 system with 24 GB of RAM, using a Virtual Box (VirtualBox 4.2.8 for Windows hosts,

www.virtualbox.org/wiki/downloads). The QIIME Virtual Box is a virtual machine based on Ubuntu Linux, which comes pre-packaged with QIIME's dependencies. Greengenes 16S alignment and Lanemask files were downloaded into QIIME prior to starting.

A tab-delimited mapping file was constructed, which contained sample-specific information required to perform the data analysis. This includes the name of each sample, the barcode sequences, the primer sequences, and any metadata information about the samples that could be used to sort the data. During the analysis, it was noted that many of the sequences were in the reverse orientation, and QIIME failed to read these sequences as the reverse primer sequence was at the 5' end instead of the 3' of the sequence. To work around this, two mapping files were created, a Forward and a Reverse file; the reverse file swapped the two primers around to allow for sequences being in the opposite orientation. The mapping files used are provided in Appendix II. Each file was checked through QIIME to ensure they were formatted correctly.

2.3.4.3.2 QIIME pipeline

Figure 2.3 shows the basic analysis steps involved in the QIIME pipeline.

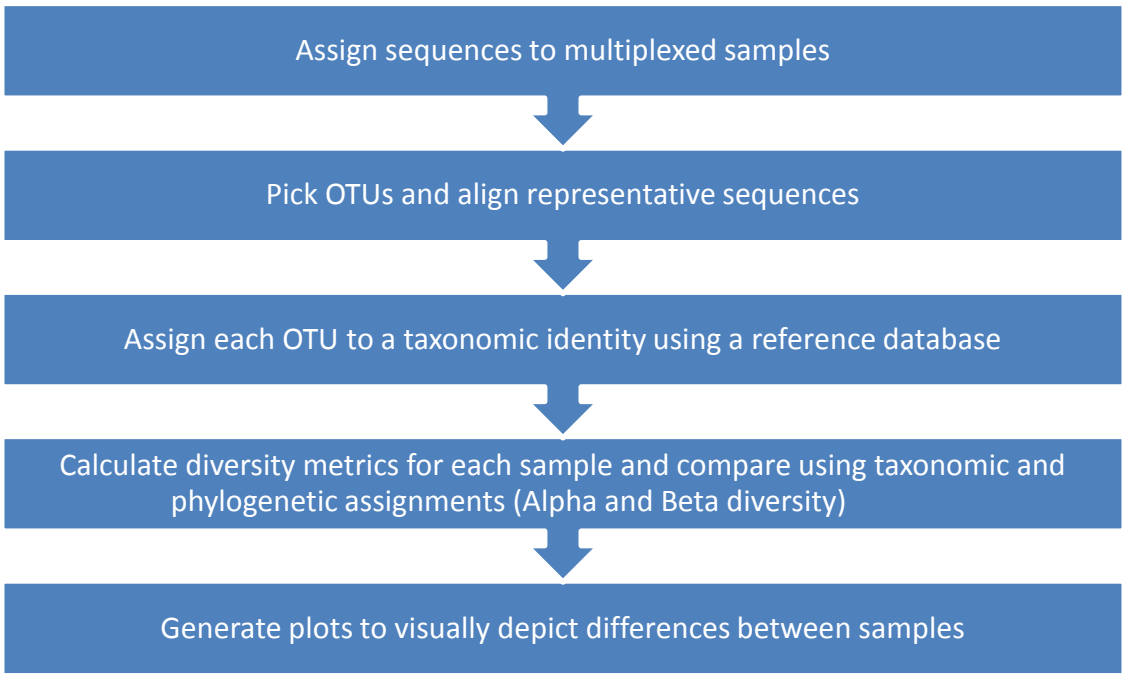


Figure 2.3: Summary of the steps involved in the QIIME pipeline.

The first step of the QIIME pipeline, "split_libraries.py", takes the raw sequencing data in FASTA format, as well as a quality file, and splits the sequences up based on their barcodes, as defined by the information in the mapping file. The raw data obtained from the University of Auckland was in a sff file format, and Geneious was initially used to remove the 454 adaptor sequences from the start of each read, with the updated file extracted as FASTA and quality files. The QIIME default parameters for filtering data were kept, with a minimum/maximum length of 200/1000, minimum quality score of 25, maximum length of homopolymers of 6, no ambiguous bases allowed and no mismatches allowed in the primer sequence. Both the forward and reverse primers were removed, as well as the barcode sequence, to ensure these sequences do not interfere with later analyses such as OTU picking and taxonomic assignments. This process was performed twice, using the two different mapping files.

The output of this command included a new FASTA formatted file where each sequence is renamed according to the sample it came from. In addition, a log file is generated, which summarises the results of data splitting, including the number of reads filtered out due to quality considerations and the number of reads assigned to each sample. The reverse sequences were reverse complemented to ensure all sequences were in the same orientation for downstream analysis, with the sequence directly following the forward primer at the 5' end, and sequence directly following the reverse primer at the 3' end. The two sequence files needed to be combined into one for downstream analysis, which was achieved using a concatenate sequences command. In order for this to successfully combine the sequences, one of the multiplex commands needed to have an extra option added in to control the starting value of the QIIME numbering system. This ensures that no sequence has the same numerical name, and was included in the reverse library script. The two aquifer samples not being analysed here were at this point removed from the total data set.

The second step of the pipeline involves a series of small steps that are performed in order automatically via the "pick_otus_through_otu_table.py" script. The workflow consists of seven steps: picking OTUs through clustering of the samples based on sequence similarity using Uclust (Edgar, 2010); selecting a representative sequence set which contains one sequence from each OTU; assigning taxonomic identities to each representative OTU sequence using the RDP classifier; alignment of the representative sequences using PyNAST (Caporaso *et al.*, 2009); filtering of the sequences to remove

gaps and excessively variable locations using the default Lanemask file; production of a Newick phylogenetic tree of the representative OTUs, required for downstream analysis, using FastTree2 (Price *et al.*, 2010); and finally making an OTU map, which is a readable matrix of the OTU abundance in each sample. This script was run using QIIME defaults, and generated an OTU table in biom format for further downstream analysis. The "summarize_taxa_through_plots.py" script generates a variety of tables and plots grouping sequences by taxonomic assignment at the different taxonomic levels. OTUs were grouped based on species information provided in the metadata file.

2.3.4.3.3 Microbial community diversity

The microbial diversity within (α-diversity) and between (β-diversity) samples can be assessed within QIIME, to describe the diversity within the study. The "alpha_rarefaction.py" script involves multiple steps being run within a single workflow, ultimately resulting in α-diversity statistics and rarefaction plots for a number of diversity metrics. This script requires the OTU table created during the OTU workflow and the mapping file to define the sample categories, as well as the phylogenetic tree created as part of the OTU workflow if phylogenetic metrics are included. The default settings include the Chao1 index for qualitative species richness, Observed species which is a measure of unique OTUs in each sample, and Phylogenetic distance, which is divergence based. The Shannon Index, a quantitative species richness metric, was added to the list of metrics calculated by creating a custom parameters file.

β-diversity, the comparison of different samples based on microbial community composition, is constructed using the "beta_diversity_through_plots.py" script, which also combines a number of steps into a single workflow, resulting in PCoA plots for each beta diversity metric selected for. The default settings were used, consisting of weighted and unweighted UniFrac phylogenetic measures (Lozupone and Knight, 2005; Lozupone *et al.*, 2007). As for the α-diversity workflow, the OTU table, mapping file and phylogenetic tree are all required.

The "jackknifed_beta_diversity.py" workflow estimates the uncertainty in the PCoA plots and hierarchical clustering through jackknife replicates, and follows a similar workflow to the one used for computing β diversity. It utilises Unweighted Pair Group Method with Arithmetic mean (UPGMA) to cluster samples using average linkage. A

bootstrapped tree is generated for both the weighted and unweighted UniFrac data, which shows how well supported the sample clustering is.

2.4 Results

2.4.1 Sample preparation

The protocol outlined in sections 2.3.1 and 2.3.2 was successful in amplifying the target 16S rRNA sequence, with the resulting samples sent for sequencing (Figure 2.4). A raw data sff file was provided by Auckland University for each of two sequencing runs performed on the same amplicon samples, GS454-01A and GS454-01B. This study utilises the small data set provided by the GS454-01A sequencing run as a proof-of-concept dataset. The GS454-01A raw data file contained 10,409 sequences, approximately one tenth of what would normally be expected from a GS 454 Junior platform.

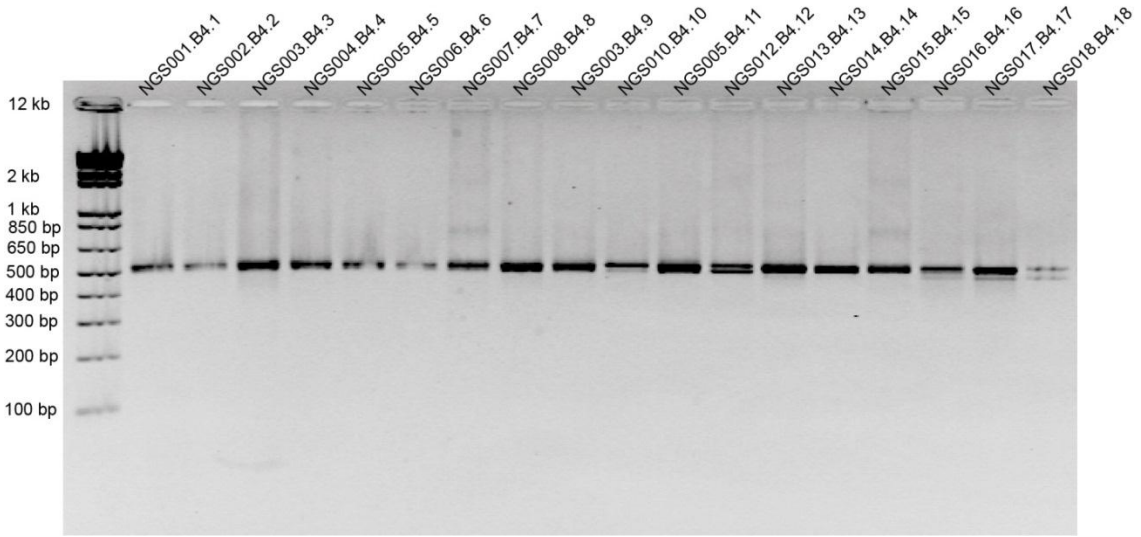


Figure 2.4: Agarose gel of final amplicon samples for GS454-01 sequencing. All samples are at a concentration of 5 ng/µl. Lane one contains 1 kb plus DNA ladder (Invitrogen).

2.4.2 Data analysis programmes

2.4.2.1 Geneious

The results from the filtering steps carried out on the raw data are summarised in Table 2.5. The filtering produced 18 sequence libraries, with read numbers ranging from 140 to 1,052 sequences.

2.4.2.2 RDP Classifier

The hierarchy file was used to analyse the community composition of each sample, and bar graphs were produced for comparison at different classification levels (Figures 2.5 and 2.6). All library species samples were combined to give a total view of bacteria present, while the three water samples were kept as individual samples.

Table 2.5: Initial processing and filtering steps of GS454-01A data in Geneious.

	Read numbers
Initial number of sequences	10409
Sequences with 454 adaptor	10197
Sequences with correct barcode	9427
Sequences at least 200 bp long	9000
Sequences with correct primer(s) sequences	8737
NGS001-B4.1	556
NGS002-B4.2	273
NGS003-B4.3	1052
NGS003-B4.9	355
NGS004-B4.4	883
NGS005-B4.5	265
NGS005-B4.11	358
NGS006-B4.6	312
NGS007-B4.7	838
NGS008-B4.8	369
NGS010-B4.10	274
NGS012-B4.12	639
NGS013-B4.13	616
NGS014-B4.14	500
NGS015-B4.15	670
NGS016-B4.16	164
NGS017-B4.17	171
NGS018-B4.18	140
Aquifer sequences included in run (not analysed)	302

2.4.2.3 QIIME

2.4.2.3.1 Data filtering and OTU selection

The data from the filtering steps provided by the "split_libraries.py" script are summarised in Table 2.6. This step was performed twice to ensure reads which had been sequenced from both the forward and the reverse primers were included in the final analysis. The sequences were concatenated into one file for OTU selection (Table

2.7). This data were imported into Microsoft Excel to produce graphs similar to those created with the RDP data (Figures 2.5 and 2.6).

Table 2.6: QIIME data from "split_libraries.py" script. This script includes filtering and splitting of sequences based on barcode.

	GS454-01A Forward map	**GS454-01A Reverse map**
Raw input sequences	10197	10197
Failed size check	327 (3.2%)	327 (3.2%)
Failed ambiguous bases	115 (1.1%)	115 (1.1%)
Failed mean quality score	738 (7.2%)	738 (7.2%)
Failed homopolymers	21 (0.2%)	21 (0.2%)
No primer match	6446 (63.2%)	3511 (34.4%)
Total sequences written to file	2550 (25.0%)	5482 (53.8%)
Total concatenated sequences	8032 (78.8%)	8032 (78.8%)
Minimum no. sequences/sample	14	22
Maximum no. sequences/sample	341	674
NGS001-B4.1	110	357
NGS002-B4.2	60	198
NGS003-B4.3	341	674
NGS003-B4.9	87	238
NGS004-B4.4	223	582
NGS005-B4.5	115	117
NGS005-B4.11	132	178
NGS006-B4.6	51	225
NGS007-B4.7	161	600
NGS008-B4.8	82	231
NGS010-B4.10	88	152
NGS012-B4.12	224	420
NGS013-B4.13	209	379
NGS014-B4.14	151	326
NGS015-B4.15	208	421
NGS016-B4.16	64	96
NGS017-B4.17	72	92
NGS018-B4.18	61	55
Aquifer sequences included in run (not analysed)	111	141

2.4.2.3.2 Microbial community diversity

Rarefaction plots for the averages of each α-diversity metric for the different sources are shown in Figure 2.7. No error bars are included for the duck and swan samples, as these sources contained only one sample. A sampling depth value, -e, was included in the parameters to provide even sampling across all samples, with the smallest number of OTU sequences found in a sample used (Table 2.7).

Two-dimensional and three-dimensional plots are generated as part of β-diversity workflow; Figure 2.8 depicts the two-dimensional beta diversity plots generated for both weighed and unweighted UniFrac measures for continuous and discrete analysis. No difference is observed between the plots for continuous verses discrete analysis.

A bootstrapped tree was generated for both the weighted and unweighted UniFrac data, which provides support for the β-diversity sample clustering (Figure 2.9).

Table 2.7: Summary of OTU data generated through QIIME using the "pick_otus_through_otu_table.py" script.

	GS454-01A
Total OTUs	2846
Minimum sequences per sample	116
Maximum sequences per sample	1015
Mean	432.2
S.D.	249.02
Suggested –e value	116
NGS001-B4.1	467
NGS002-B4.2	258
NGS003-B4.3	1015
NGS003-B4.9	325
NGS004-B4.4	805
NGS005-B4.5	232
NGS005-B4.11	310
NGS006-B4.6	276
NGS007-B4.7	761
NGS008-B4.8	313
NGS010-B4.10	240
NGS012-B4.12	644
NGS013-B4.13	588
NGS014-B4.14	477
NGS015-B4.15	629
NGS016-B4.16	160
NGS017-B4.17	164
NGS018-B4.18	116

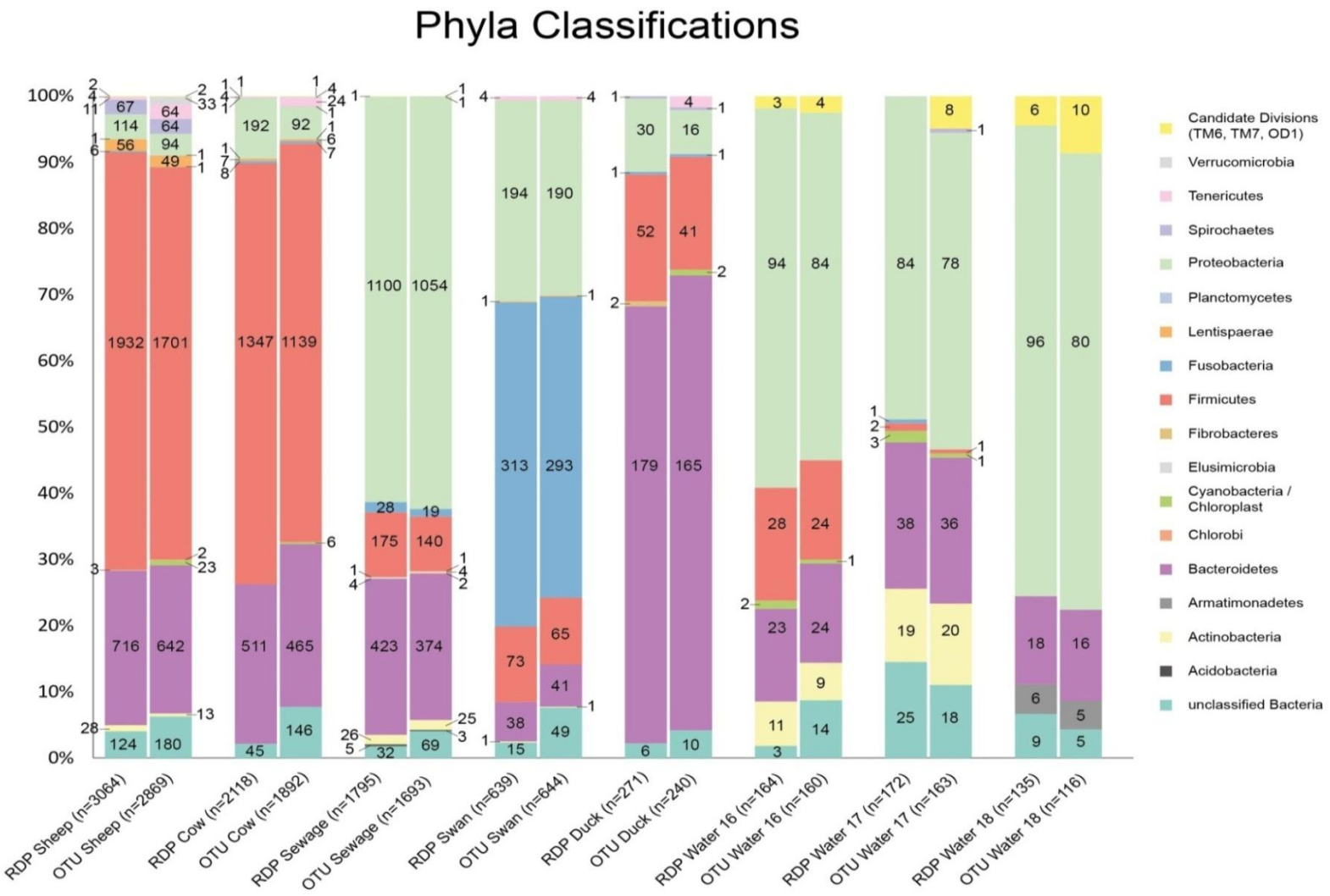


Figure 2.5: Phyla taxonomy level classifications for faecal source library and water samples. The left hand column for each library and water sample shows the sequence phyla classifications using the RDP Classifier; the right hand column shows the OTU classifications through the QIIME platform. Numbers in each bar represent the number of OTUs assigned to the corresponding phylum**.**

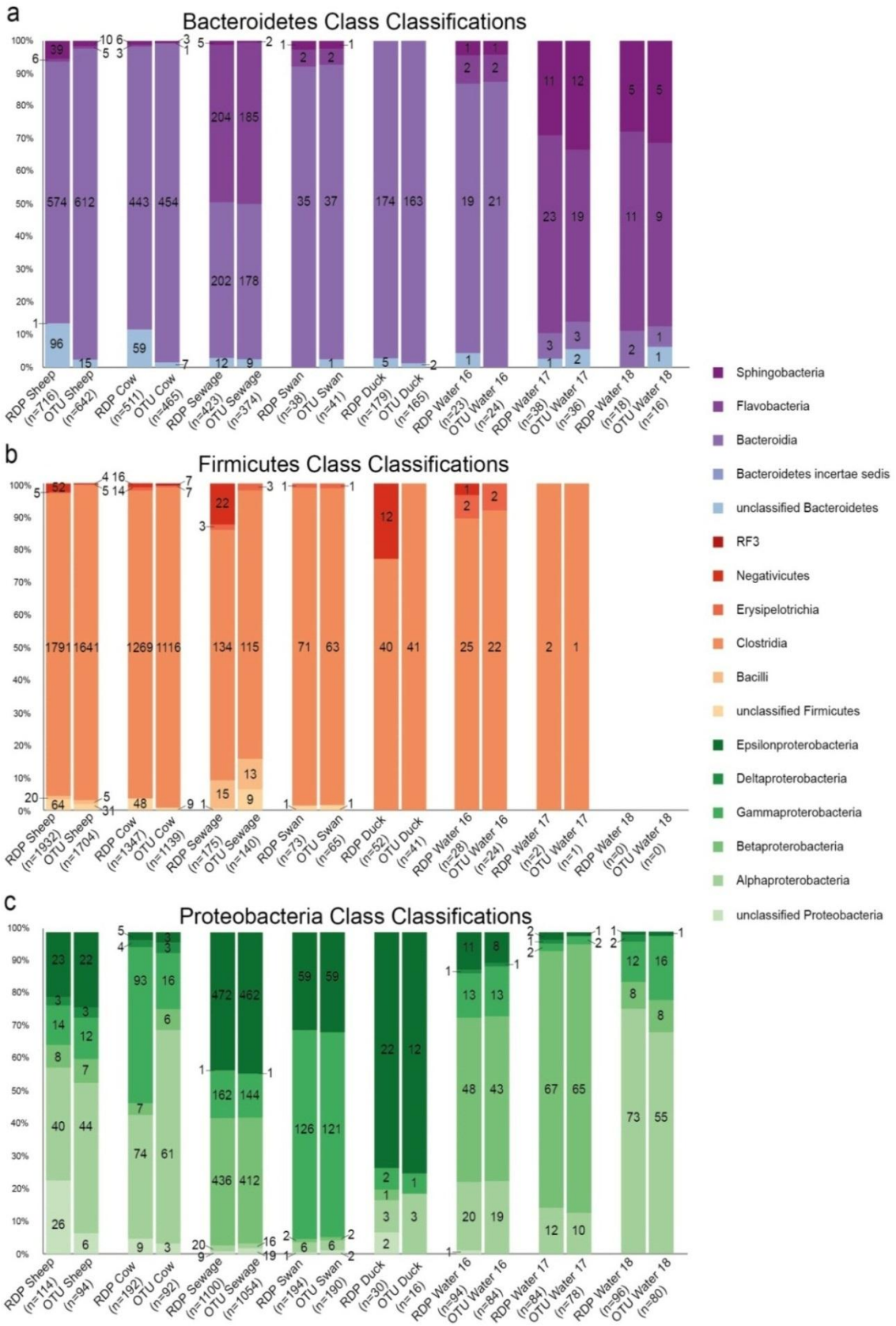


Figure 2.6: Class taxonomy level classifications for the three phyla found throughout the faecal source library and water samples; Bacteroidetes (a), Firmicutes (b) and Proteobacteria (c). The left hand column for each library and water sample shows the sequence phyla classifications using the RDP Classifier; the right hand column shows the OTU classifications through the QIIME platform. Numbers in each bar represent the number of OTUs assigned to the corresponding phylum.

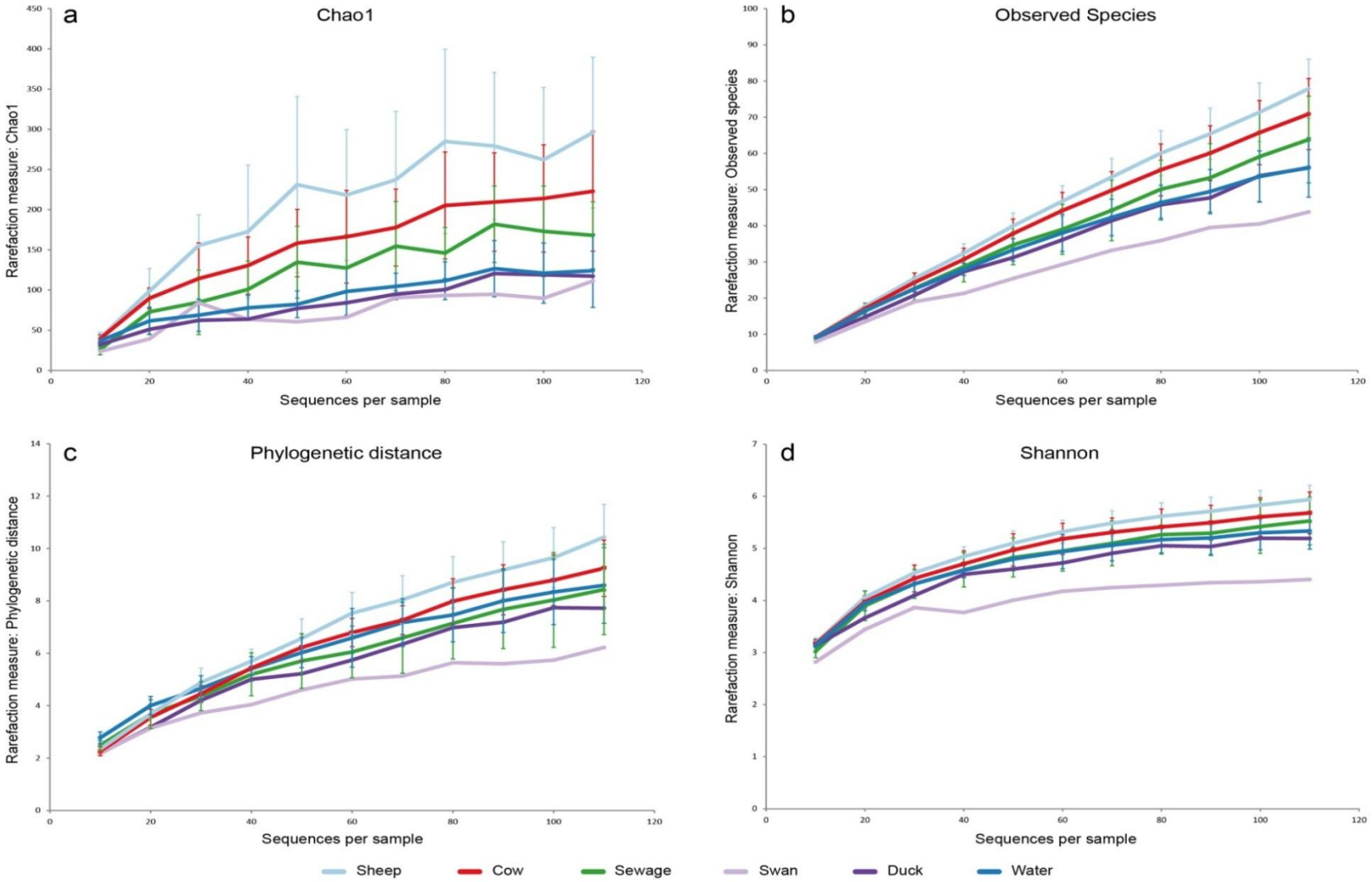


Figure 2.7: Rarefaction curves produced by QIIME using four different alpha diversity metrics: Chao1 (a), Observed species (b), Phylogenetic Distance (c) and Shannon (d). For species where more than one sample was analysed, the average is plotted with error bars displaying the sample variation.

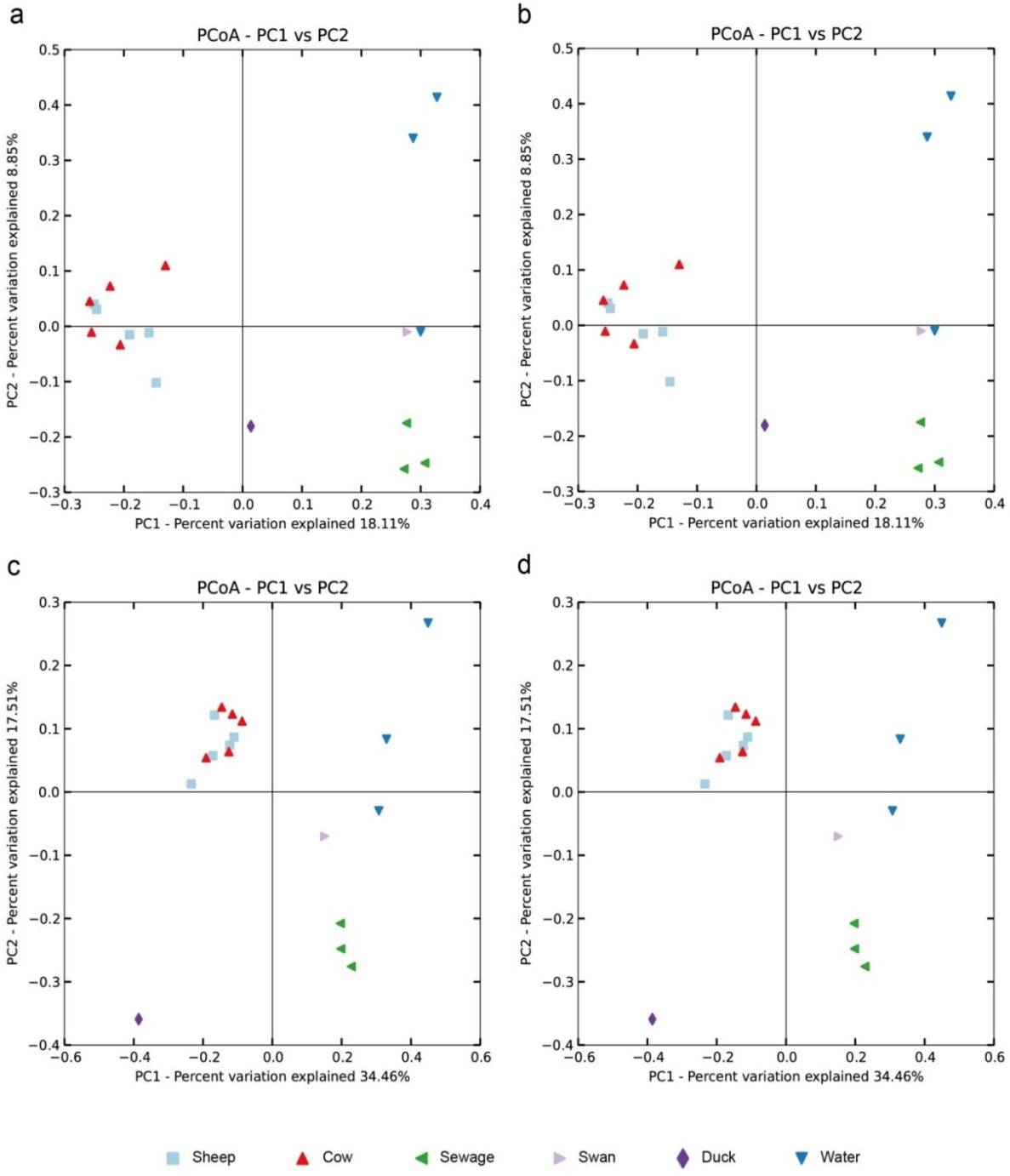


Figure 2.8: Two-dimensional PCoA UniFrac plots generated by QIIME; unweighted continuous (a), unweighted discrete (b), weighted continuous (c) and weighted discrete (d).

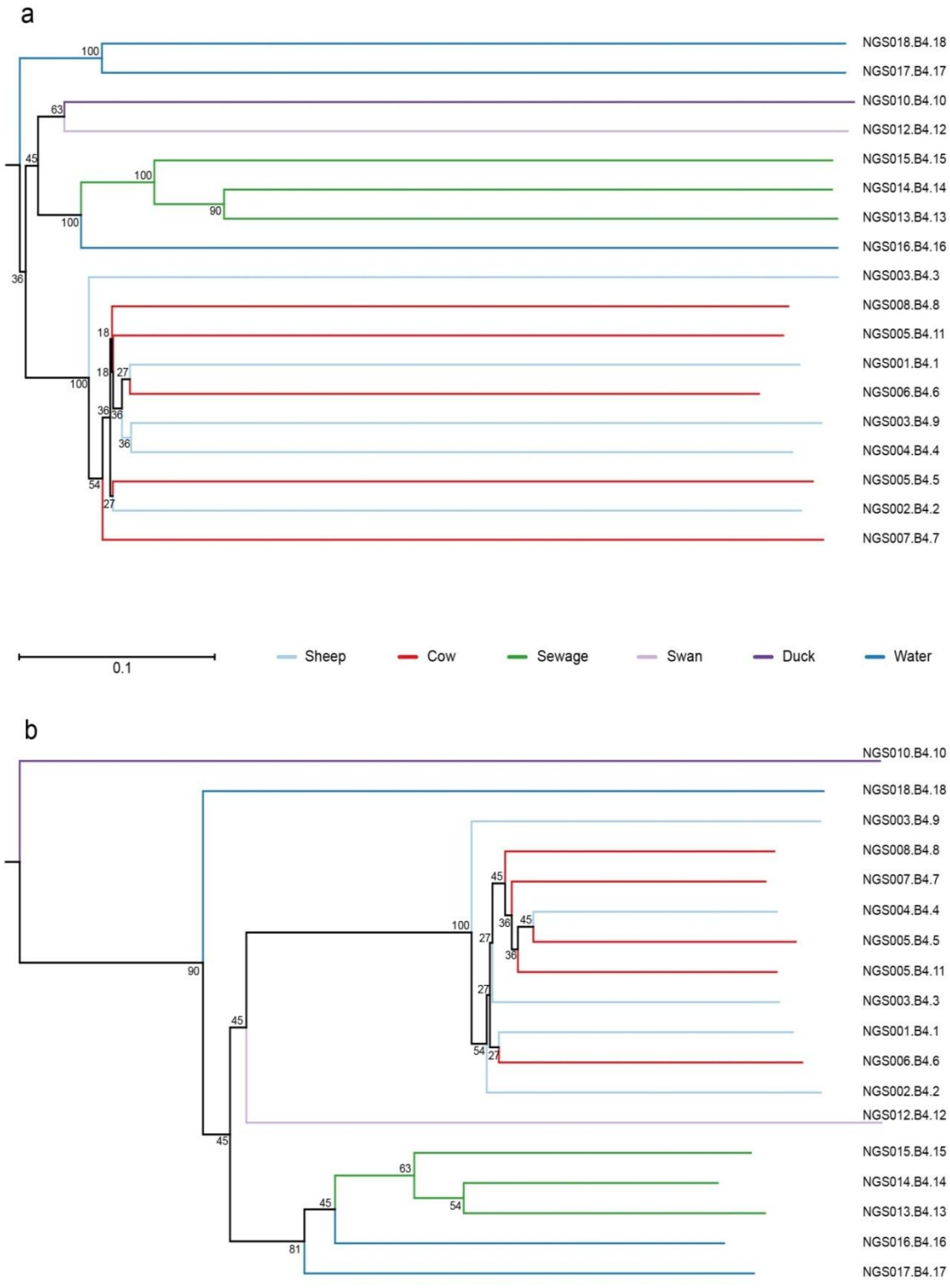


Figure 2.9: Jackknifed UPGMA bootstrapped trees. Unweighted UniFrac data clustering (a) and weighted UniFrac data clustering (b).

2.5 Discussion

2.5.1 Sample preparation

The majority of the DNA sequences were able to be taxonomically classified using the RDP Classifier (Figure 2.5). This suggests that the V1-V3 hypervariable regions of the 16S rRNA gene amplified and sequenced were suitable for identification of the different bacteria present, with less than 15% not assigned to a bacterial phylum for any one sample. A number of the sequencing reads contained both the forward and reverse primer sequences, suggesting that the target amplicon length is suitable for sequencing using the 454 platforms, and could possibly be extended to include the V4 region as well.

2.5.1.1 Amplification protocols

There were notable differences between the numbers of sequences generated for the sheep sample that was included twice, using two different amplification protocols. NGS003.B4.9 used the same protocol as all the other samples, and resulted in 355 sequences, whereas NGS003.B4.3 used an amplification protocol without a specific annealing step. This sample contained 1,052 sequences, which was by far the largest number of sequences generated for any one sample (Table 2.5). The cow sample which was also included twice with the same two protocols showed very similar sequencing numbers, with 358 reads for the standard protocol and 265 reads for the two-temperature protocol. Figure 2.4 shows similar sized bands for all four samples, suggesting that the large difference in read numbers seen in the two sheep samples is most likely due to sequencing. As the problems surrounding the final library preparation step are unknown, this ligation step may have also influenced the number of sequences generated for each sample. When the individual taxonomic classifications of these samples are considered (data not shown), the percentages of sequences assigned to each phyla were generally similar.

In the sheep samples, only two phyla (*Actinobacteria* and *Fusobacteria*) were included in the NGS003.B4.3 sample that were not in the NGS003.B4.9 sample, but only accounted for 1.4% of the total sequences for that sample. For the cow samples, three phyla (*Fusobacteria*, *Lentisphaerae* and *Tenericutes*) were in the NGS005.B4.11 sample but were not in the NGS005.B4.5 sample, accounting for 1.7%. That the

differences between the two protocols were not consistent suggests that there are no measureable differences between the two protocols, and highlights the need to ensure the same amplification protocol is used for all samples, enabling accurate comparisons.

2.5.2 Data analysis programmes

2.5.2.1 Taxonomy classifications

Both analysis systems used were able to classify the majority of the sequences based on the 16S rRNA database provided by the RDP Project, with an average of 3.1% and 6.3% for sequences not classified to bacterial phyla for the RDP and QIIME analyses, respectively. The QIIME platform had a higher portion not assigned due to a number of reasons. The default confidence level for the QIIME "assign_taxonomy.py" script is 0.97, whereas this was set to 0.6 for the RDP Classifier, as a 0.5 level has been shown to accurately classify partial 16S rRNA sequences shorter than 250 bp (Claesson *et al.*, 2009). The QIIME platform also includes a more stringent filtering system than what was used within Geneious prior to loading sequences into the RDP Classifier, with 8,435 and 7,780 sequences after filtering for the RDP Classifier and QIIME, respectively.

Figure 2.5 provides the comparison of the taxonomy classifications made by both the RDP Classifier and the QIIME platforms. The term 'operational taxonomic units' (OTUs) has generally substituted the concept of 'species' in metagenomics, and defines the sequence similarity at a given threshold as belonging to the same taxonomic level (Kuczynski *et al.*, 2012; Zarraonaindia *et al.*, 2013). Overall, the two sets of classifications match up reasonably well, with similar percentages for each of the phyla. The QIIME OTU results generally show slightly more unclassified Bacteria, and slightly less in some of the phyla, as noted above. The biggest discrepancy between the two classifications is in sheep sequences, where QIIME has classified 23 sequences as *Cyanobacteria*/Chloroplast, whereas the RDP Classifier has assigned none. QIIME has also found a few of these sequences in the cow and duck libraries as well as in the water samples, while the RDP Classifier has only identified them in the water samples. However, as these sequences are most likely to be environmental, they are not likely targets for MST. The QIIME analysis has picked up a couple of extra phyla, including *Chlorobi*, which are green sulfur bacteria, and a couple of Candidate divisions, *OD1*

and *TM6*. These were only identified in low numbers and are not likely MST targets (refer to Section 1.1.3).

Bacteroidetes, *Firmicutes* and *Proteobacteria* were dominant across the majority of the faecal libraries. The exception to this was the swan sample, which had a very high percentage of *Fusobacteria* sequences. This is comparable to recent studies that looked at faecal samples from a range of sources (Jeong *et al.*, 2011; Ley *et al.*, 2008a; Unno *et al.*, 2010). However, the percentage of *Proteobacteria* in the human sewage library, approximately 60%, is much higher than reported elsewhere for human samples. This may be due to the human-specific samples used in this study being DNA extracted from raw sewage rather than individual faecal material, although a recent study looking at the microbial population of sewage from a wastewater plant found only 25% of the population to be *Proteobacteria* (McLellan *et al.*, 2010). Amplification bias may explain these differences, due to different amplicon primers being used for each study.

Figure 2.6 depicts the next level of taxonomic classification for these three phyla, which shows there is a reasonable level of variation in the classes of *Bacteroidetes* and *Proteobacteria* across the different source libraries, and the majority of the *Firmicutes* belonging to the *Clostridia* class. As this was a proof-of-concept study, the analyses were not targeted to source-specific or novel sequences, and further work would allow more taxonomic classifications and comparisons to determine if there are any sequences which are source-specific.

2.5.2.2 Diversity measures

QIIME has the advantage of including α- and β-diversity measures, which are important parameters for analysing communities (Lozupone *et al.*, 2007). The α-diversity includes rarefaction plots, which compare the number of OTU sequences as a function of individuals sampled. The plot usually starts as a steep slope, which flattens out over time as fewer species are discovered per sample (Wooley *et al.*, 2010). There are a number of different statistical analyses that can be used to estimate species diversity, each with their own advantages and disadvantages (Lozupone and Knight, 2008), and depending on which one is used, the results can differ quite markedly, as can be seen in Figure 2.7. Based on the curve generated by the Observed species metric, there is still a large amount of diversity unlikely to be sampled, as the plots for all samples are still trending upwards. In comparison, the plots for the Shannon index, a species based

quantitative metric, show that the diversity is mostly captured by the current samples, as the plots are all levelling out.

The β-diversity analysis includes PCoA, one of the most commonly used dimensionality reduction techniques in microbial ecology (Gonzalez and Knight, 2012), utilising the UniFrac metric, which measures the phylogenetic distance between sets of taxa in a phylogenetic tree as a fraction of the branch length of the tree that leads to descendants from either one of two communities (Lozupone and Knight, 2005). A relatively small UniFrac distance implies that the two communities are similar, with lineages that share a common evolutionary history (Costello *et al.*, 2009). Figure 2.8 shows the clustering of each sample based on UniFrac measurements. The sheep (light blue) and cow (red) samples cluster very closely together, implying that the microbial communities within these two species are almost identical. The human sewage samples (green) cluster together quite separate from any of the other samples, as does the duck (purple) sample. The water samples are reasonably spread out from each other, and do not cluster close to any of the other samples, with the exception of the swan (light purple) sample for the unweighted UniFrac analyses (a and b), which clusters closely to one of the water samples. This highlights the taxonomic diversity between the three water samples seen in Figure 2.5, which could be expected as each water sample is from a different part of New Zealand and would be likely to have different levels of impact. The phylogenetic trees (Figure 2.9) provide support for the β-diversity clustering, with bootstrap values indicating the confidence of support for each cluster node. While the unweighted tree (a) has some high node values, it also contains quite a few below 25, suggesting less than 25% support for those clusters; however, the weighted tree (b) has more nodes with greater than 25% support.

Based on data generated so far, we are unable to make a recommendation on whether any of the water samples have been contaminated with faecal material. A larger sample set would be beneficial to determine if the variation seen across different sources is due to individual differences or source-specific variation. The limited sequencing reads in this data set may also impact the results, and a greater sequencing depth for these samples is recommended. A greater understanding on what influences the clustering of faecal samples may provide further insight into ensuring water samples cluster according to contamination source.

2.5.3 Limitations of the methods

2.5.3.1 Barcoding

During the initial processing of the sequences from the GS454-01A data, insertions within the primer and barcode sequences were apparent in a number of sequences. If an insertion or deletion were to occur in the right place, the four nt barcode could be altered such that the expected barcode position then matches another barcode sequence, causing the sequence to be classified as coming from a different source. By using stringent filtering for only completely matching 454 adaptor and barcode sequences, the confidence of correct sample assignment by barcode for these samples is fairly high; however, the use of longer barcode sequences would provide greater confidence that no sequencing artefacts have altered the barcode. Roche supply a 454 sequencing barcode kit (Multiplex Identifiers, MIDs), which provides 12 barcoding reactions, including the 454 sequencing adaptors, to add a 10 nt barcode to each library, however, this option would be expensive for the number of samples in this study needing to be uniquely identified. Roche also provide a list of over 150 MIDs they had developed, which can be incorporated into primer design. Because the length of sequencing reads obtained through 454 sequencing is only approximately 500 base pairs, the length of barcodes used can have a large effect on the final length of the usable sequence obtained. For this reason, it has been suggested that the use of 10 nt long barcodes is not necessary, and six or eight nucleotides would be a better compromise between confidence in correctly assigning sequences to their source and getting the most usable sequence data possible.

Binary coding schemes, such as Hamming coding have been used to create error-correcting barcode sequences for use in next generation sequencing applications (Bystrykh, 2012; Hamady *et al.*, 2008). These barcodes utilise 2 bit binary words to encode each nucleotide symbol; using an alphabetical order, A will be encoded as 00, C as 01, G as 10 and T as 11. The Hamming binary codes are translated into a nucleotide sequence by converting every two consecutive bits into the DNA nucleotide code (Bystrykh, 2012). This system also results in an ability to detect and correct errors in the binary format. Using the Hamming 8,4 sequences provided by Bystrykh (2012) allows the use of slightly longer 8 nt barcodes as well as providing an intrinsic ability to detect substitution errors that may occur during the sequencing process.

2.5.3.2 PCR design

A limitation of the way the amplicons were prepared is that the amplicon library still requires an additional ligation step where the 454 sequencing adaptors are added to the 5' end of each sequence. This step can be removed if the PCR design incorporates the addition of the sequencing adaptors. Roche recommend the use of fusion primers, which incorporate the 454 sequencing adaptor sequences, the barcode sequence and the target sequence all at once, resulting in primers approximately 60 nt long. A recent paper on the use of next generation sequencing for multilocus sequence typing (Boers *et al.*, 2012) designed a two-step PCR method, utilising universal 'tail' sequences that allow barcodes and sequencing adaptors to be added to each sample during a second PCR, without having to incorporate target-specific sequences for each fusion primer (Figure 2.10). This results in a much more flexible and cost-effective way of preparing samples for next generation sequencing, particularly when using multiple target-specific primers within one sample.

This method was trialled using the same Bac8F and Univ529R primer sequences, each with a specific universal tail added (Table 2.8), following the same reaction mix as outlined in Table 2.4. The sequences were cleaned using the AMPure XP beads protocol outlined in section 2.3.2.3. The product from this PCR was then used as the template for the second PCR, with fusion primers incorporating a unique barcode sequence (Bystrykh, 2012), the 454 sequencing adaptor, and the universal tail added to the 5' end of each primer in the first PCR being the target sequence (Table 2.8). The PCR protocol used was modified from Boers *et al.* (2012): initial denaturation at 95^{o}C for 2 min followed by 35 cycles with cycling conditions of 30 s at 95^{o}C, 30 s at 50^{o}C and 60 s at 68^{o}C. During the first 10 cycles, the annealing temperature was increased by 0.5^{o}C per cycle to an annealing temperature of 55^{o}C. A final extension followed for 2 min at 68^{o}C. The first PCR step worked well for the samples from this study, but optimisation of the second PCR proved difficult and amplification of samples was inconsistent.

Discussions with people who regularly run 454 sequencing suggested that the optimisation required to get strong amplification results using fusion primers was often difficult and hard to obtain. Recommendations were to continue the use of the ligation step after the PCR library had been prepared.

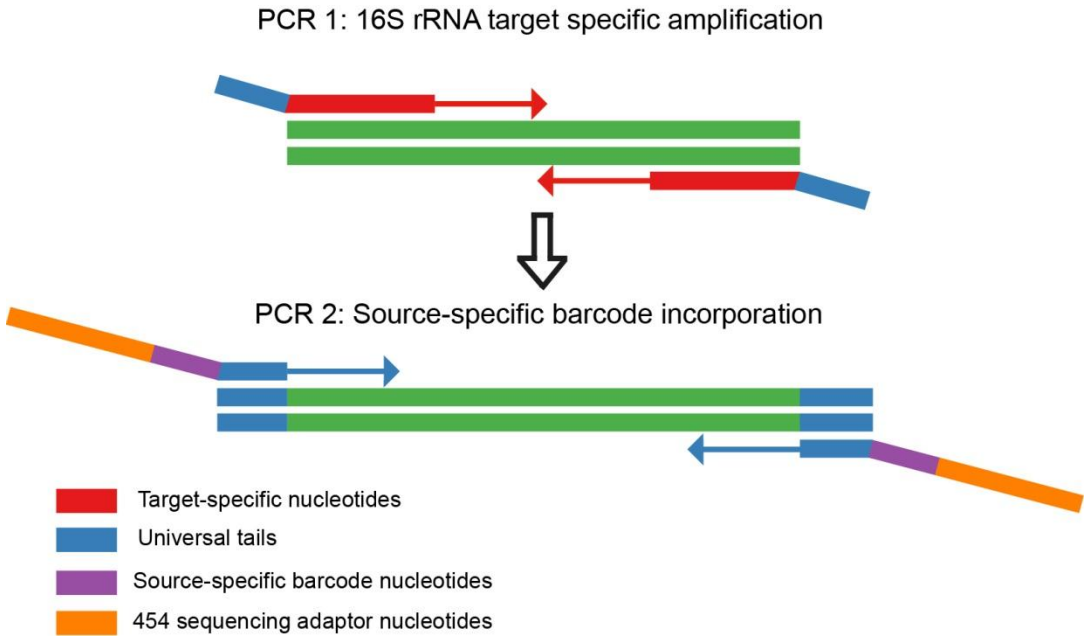


Figure 2.10: Two-step PCR protocol; adapted from Boers *et al.* (2012).

Table 2.8: PCR primers used for trialling the two-step amplification method. Barcode sequences are highlighted in bold.

Primer	Nucleotide sequence	T_m (°C)
Tailed Bac8F	**GACACTATAG**AGAGTTTGATCCTGGCTCAG	57
Tailed Univ529R	**CACTATAGGG**ACCGCGGCKGCTGGC	62
454FA-H1-tailA	CCATCTCATCCCTGCGTGTCTCCGACTCAG**AGAGAGAG**GACACTATAG	68
454RB-H1-tailB	CCTATCCCCTGTGTGCCTTGGCAGTCTCAG**AGAGAGAG**CACTATAGGG	69
454FA-H2-tailA	CCATCTCATCCCTGCGTGTCTCCGACTCAG**TCACAGCA**GACACTATAG	68
454RB-H2-tailB	CCTATCCCCTGTGTGCCTTGGCAGTCTCAG**TCACAGCA**CACTATAGGG	69
454FA-H3-tailA	CCATCTCATCCCTGCGTGTCTCCGACTCAG**GTAGCACT**GACACTATAG	68
454RB-H3-tailB	CCTATCCCCTGTGTGCCTTGGCAGTCTCAG**GTAGCACT**CACTATAGGG	69
454FA-H4-tailA	CCATCTCATCCCTGCGTGTCTCCGACTCAG**ATAGCGTC**GACACTATAG	68
454RB-H4-tailB	CCTATCCCCTGTGTGCCTTGGCAGTCTCAG**ATAGCGTC**CACTATAGGG	69
454FA-H5-tailA	CCATCTCATCCCTGCGTGTCTCCGACTCAG**CTAGCTGC**GACACTATAG	68
454RB-H5-tailB	CCTATCCCCTGTGTGCCTTGGCAGTCTCAG**CTAGCTGC**CACTATAGGG	69

Table 2.8 continued

Primer	Nucleotide sequence	T_m (°C)
454FA-H6-tailA	CCATCTCATCCCTGCGTGTCTCCGACTCAG**CTACGACA**GACACTATAG	68
454RB-H6-tailB	CCTATCCCCTGTGTGCCTTGGCAGTCTCAG**CTACGACA**CACTATAGGG	69
454FA-H7-tailA	CCATCTCATCCCTGCGTGTCTCCGACTCAG**GTACGCAT**GACACTATAG	68
454RB-H7-tailB	CCTATCCCCTGTGTGCCTTGGCAGTCTCAG**GTACGCAT**CACTATAGGG	69
454FA-H8-tailA	CCATCTCATCCCTGCGTGTCTCCGACTCAG**ACATGCGT**GACACTATAG	68
454RB-H8-tailB	CCTATCCCCTGTGTGCCTTGGCAGTCTCAG**ACATGCGT**CACTATAGGG	69
454FA-H9-tailA	CCATCTCATCCCTGCGTGTCTCCGACTCAG**GCATGTAC**GACACTATAG	68
454RB-H9-tailB	CCTATCCCCTGTGTGCCTTGGCAGTCTCAG**GCATGTAC**CACTATAGGG	69
454FA-H10-tailA	CCATCTCATCCCTGCGTGTCTCCGACTCAG**ATACGTGC**GACACTATAG	68
454RB-H10-tailB	CCTATCCCCTGTGTGCCTTGGCAGTCTCAG**ATACGTGC**CACTATAGGG	69
454FA-H11-tailA	CCATCTCATCCCTGCGTGTCTCCGACTCAG**GCAGTATC**GACACTATAG	68
454RB-H11-tailB	CCTATCCCCTGTGTGCCTTGGCAGTCTCAG**GCAGTATC**CACTATAGGG	69
454FA-H12-tailA	CCATCTCATCCCTGCGTGTCTCCGACTCAG**TCAGTCGA**GACACTATAG	68
454RB-H12-tailB	CCTATCCCCTGTGTGCCTTGGCAGTCTCAG**TCAGTCGA**CACTATAGGG	69
454FA-H13-tailA	CCATCTCATCCCTGCGTGTCTCCGACTCAG**ACAGTGCT**GACACTATAG	68
454RB-H13-tailB	CCTATCCCCTGTGTGCCTTGGCAGTCTCAG**ACAGTGCT**CACTATAGGG	69
454FA-H14-tailA	CCATCTCATCCCTGCGTGTCTCCGACTCAG**TGCTACAG**GACACTATAG	68
454RB-H14-tailB	CCTATCCCCTGTGTGCCTTGGCAGTCTCAG**TGCTACAG**CACTATAGGG	69
454FA-H15-tailA	CCATCTCATCCCTGCGTGTCTCCGACTCAG**AGCTAGTC**GACACTATAG	68
454RB-H15-tailB	CCTATCCCCTGTGTGCCTTGGCAGTCTCAG**AGCTAGTC**CACTATAGGG	69

2.5.3.3 Data analysis programmes

The different programmes trialled using this data all proved difficult to use on their own to fully analyse the data, suggesting the best way to process this data is by using a suite of programmes.

Geneious proved useful for controlling the initial filtering of data, especially for removing the sequencing adaptors from the sequences prior to loading into another programme. Geneious also provides various export format options, including quality files which are often needed for the other programmes. However, Geneious does not appear to have many post-processing data analysis options included, and what it does do takes a large length of time.

The RDP project website proved easy to use, although uploading data could sometimes take a lengthy amount of time, depending on the number of users on the site at any one given point of time. The website also occasionally dropped out through the middle of analysis. The data generated by the RDP Classifier was easy to understand and appeared to classify a reasonable number of the sequences. However, the RDP database does not classify down to the species level, which may be required if this data were to be used to design source-specific markers. Other databases are available, such as SILVA (Quast *et al.*, 2013), which do contain sequences classified to species level, but do not have the online classification ability that RDP does, so must be used within a programme such as QIIME. As QIIME is pre-packaged with the RDP as the classifier, to change the database would require extra knowledge on how to incorporate other packages.

QIIME provided a simple workflow scheme for working through data relatively fast and easily, with little background knowledge needed. With the help provided from the QIIME website and tutorials, managing the scripts needed to run the programme is straightforward, although because it is a Linux command-based system, it can be daunting at first for those not used to this style of platform. Troubleshooting issues with a script can be difficult if you do not fully understand what is happening during the workflow, and can be time consuming working through all the potential faults with the script and/or input data. The results generated by QIIME are meaningful and easily interpreted, however, there is a large amount generated and the user does need to determine what is relevant and what is not. QIIME took approximately 1 day to fully run through the suggested scripts required, once the system was fully up and running

and we understood what was needed. This is an acceptable time period given the size of the dataset, and suggests QIIME would be suitable to continue to use for this study.

2.6 Conclusions

The sample preparation steps outlined are standard DNA extraction protocols and did not prove difficult or introduce downstream problems; therefore, no changes to these protocols are required. The primers selected produced an amplicon large enough to classify bacteria using the RDP Classifier, and changing the barcode region from four to eight nucleotides will assist in a higher confidence of correct assignment of sequences to their source. The Bac8F and Univ529R primers should continue to be used, but incorporate the eight-nucleotide Hamming-based barcodes published by Bystrykh (2012) at the 5' end. As the addition of the 454 sequencing adaptors through the use of fusion primers proved difficult, the use of ligation to add the adaptors to each sequence should continue to be used. This step can be completed by the sequencing provider on request.

The data analysis for this dataset required initial manipulation using Geneious, to ensure data is a good enough quality and provide the ability to remove extra sequencing tags if not already removed before further processing. It was also needed to convert the provided raw data files into the right format for QIIME; however, this may not be required depending on what raw data files are supplied by the sequencing provider. The QIIME programme offers a range of analyses that are easy to implement, and the classification output is very similar to that produced by the RDP Classifier based on the sequences directly. QIIME also provides the ability to analyse α- and β-diversity. While QIIME does require some understanding of the analyses in order to implement the ones suitable for the data, there is a large amount of information available on each script and how to use it. The types of analyses useful for 16S rRNA sequence data are also well documented.

These methods proved successful for identifying the bacterial species present in both faecal and water samples, suggesting that next generation sequencing methods support the aims of microbial source tracking. The use of readily available analysis programmes, such as QIIME, simplifies the analysis process. More understanding of how the diversity analyses cluster samples together may be needed to accurately assign

faecal source contamination to a water sample. To further this study, additional faecal samples need to be analysed using the community diversity metrics provided by QIIME, preferably with larger sequence numbers per sample than analysed in this study.

Chapter Three
Metagenomic analysis of faecal and water samples for microbial source tracking using next generation sequencing

3.1 Abstract

Advances in DNA sequencing allow for comprehensive analysis of microbial communities, particularly through the use of 16S rRNA gene profiling. The ability to barcode samples allows for the analysis of large datasets, enabling large metagenomic analysis of microbial communities from multiple environments simultaneously. This is a promising technique for the field of microbial source tracking, which aims to determine the source of faecal contamination in water bodies. 454 pyrosequencing targeting the V1-V3 hypervariable regions of the 16S rRNA gene was used to investigate the structure of bacterial communities in faecal sources and water samples from various sources in New Zealand. 522,065 raw sequencing reads were generated through three sequencing runs. 39,112 OTUs were identified and analysed for microbial community diversity. Rarefaction analysis suggests that ruminant faeces, including alpaca, cow and sheep, had the highest microbial diversity, with dog faeces having the lowest value, followed by human and swan faeces. Samples were clustered based on similarities in the phylogenetic lineages using principal coordinate analysis plots, and were found to cluster based on the host source. Removal of environmental bacteria was found to have little effect on the clustering. Three major phyla, *Bacteroidetes*, *Firmicutes* and *Proteobacteria* were identified in each sample, and a number of genera that could be useful as source specific markers were identified. These markers suggest possible ruminant or human contamination of six out of ten water samples.

3.2 Introduction

Faecal contaminated water bodies pose a risk to human health, due to the potential presence of pathogenic microorganisms, viruses and protozoa. It is important to be able to identify the source of contamination in order to implement suitable management strategies to minimise pollution and the impact on human health. A range of microbial source tracking (MST) methods have been developed as a means of identifying the source of contamination (Stoeckel and Harwood, 2007). Many of these are based on the identification of a host-specific marker, in particular based on the 16S rRNA gene, a ubiquitous gene which includes multiple conserved and hypervariable regions, allowing all bacteria in a sample to be amplified and taxonomically classified (Baker *et al.*, 2003; Chakravorty *et al.*, 2007).

Advances in DNA sequencing technologies have resulted in the ability to sequence complete environmental samples, where entire bacterial communities can be analysed from a large variety of sources (Shokralla *et al.*, 2012). Due to the large throughput abilities of next generation sequencing (NGS) platforms, multiple samples can be included in a single sequencing run, through the use of barcodes, which allow the original source of each sequence to be identified (Cardenas and Tiedje, 2008).

A number of studies have used NGS technologies to investigate the bacterial communities within a range of faecal sources, including human faeces (Costello *et al.*, 2009; Dethlefsen *et al.*, 2008), cattle faeces (Dowd *et al.*, 2008), sewage (McLellan *et al.*, 2010) and activated sludge of wastewater treatment plants (Sanapareddy *et al.*, 2009). Only a limited number of studies have utilised NGS technologies for MST applications, with the focus predominantly on analysis of faecal sources (Lee *et al.*, 2011). A couple of studies in South Korea have included water samples. Jeong *et al*. (2011) identified a number of potential source-specific faecal indicators, based on comparisons of taxonomic distribution between a range of faecal samples and river water. Unno *et al*. (2010) evaluated the source of faecal contamination in two river samples, based on a ratio of the number of operational taxonomic units (OTUs) shared between each faecal and water sample. Analysis of a third river sample identified a high level of environmental bacteria present, rather than faecal contamination.

Microbial diversity has also been assessed within a range of environmental and host samples through the use of diversity measures. Quantitative and qualitative alpha (α)

and beta (β) diversity measures have been used to reveal bacterial community differences for a range of diverse physical environments (Lozupone and Knight, 2007), different human body habitats (Costello *et al.*, 2009) and a range of mammalian species (Ley *et al.*, 2008a). These diversity measures allow for many different hypotheses about community structure to be tested using sequence information, and it has been suggested that these methods have the potential to reveal fundamental properties of microbial communities beyond what has been possible with species-based measures alone (Lozupone and Knight, 2008).

In this study, we investigated the microbial communities of major faecal sources in New Zealand, including chickens, cows, dogs, ducks, humans, sheep and swans, as well as a number of other sources, which may not contribute to faecal contamination to the same extent (alpaca, horse, pig, possum and pukeko). Ten contaminated water samples from around New Zealand were also included in the analysis. The V1-V3 hypervariable region of the 16S rRNA gene was amplified via targeted amplification, with the resulting amplicon sequenced using the Roche 454 pyrosequencing platform. The bacterial diversity was determined through the use of QIIME (Caporaso *et al.*, 2010), incorporating taxonomic classification and microbial diversity.

3.3 Methods and materials

3.3.1 Sample preparation

Details of faecal and water samples used this study are provided in Tables 3.1 and 3.2, respectively, and were selected using the same criteria outlined in Section 2.3.2.

Additional species samples were prepared and amplified; however, the targeted 16S rRNA amplification was not enough to be included in the sequencing sample. These were predominantly bird species, including seagull, Canada geese and swan samples, with a dog and a horse sample also not able to be included.

Two water samples were also selected for sequencing twice, to determine the sequencing effects on community composition.

Table 3.1: Faecal library samples used in the combined sequencing study. * indicates samples freshly collected and extracted for this study. Other samples were archived extracted DNA samples stored at 4°C. Composite samples containing DNA extracted from five individual samples were prepared after DNA extraction. Human sewage samples were not composited, due to already containing a mixed human faecal source. qPCR results are the threshold cycle (Cp) values for a general source qPCR assay.

ESR Sample ID	Previous qPCR results	Species	Location sample taken from	Study Sample ID	Barcode tag	Sequencing Run
CMB05176 CMB05177 CMB05178 CMB05179 CMB05180		Sheep (*Ovis aries)*	Lyttelton	NGS001	B4.1	GS454-01B
CMB05188 CMB05189 CMB05190 CMB05191 CMB05192	15.09 14.35 14.86	Sheep (*Ovis aries)*	Dunsandel	NGS002	B4.2	GS454-01B
CMB120037* CMB120038* CMB120039* CMB120040* CMB120041*	13.83 14.50 14.60 15.16 14.68	Sheep (*Ovis aries)*	Christchurch	NGS003	B4.9	GS454-01B
CMB120325* CMB120326* CMB120328* CMB120331* CMB120333*	15.03 14.74 14.61 14.76 15.75	Sheep (*Ovis aries)*	Winchmore	NGS004	B4.4	GS454-01B
Cawthron 7 Cawthron 8 Cawthron 9 Cawthron 10 Cawthron 11	20.75 20.67 20.37 21.60 19.62	Cow *(Bos primigenius)*	South Island	NGS005	B4.11	GS454-01B
CMB06648 CMB06649 CMB06650 CMB06651 CMB06680	22.4 22.0 21.9 21.0 16.7	Cow *(Bos primigenius)*	Cust	NGS006	B4.6	GS454-01B
CMB06684 CMB06685 CMB06686 CMB06687 CMB06688	17.86 16.2	Cow *(Bos primigenius)*	Lincoln	NGS007	B4.7	GS454-01B
MB1004001 MB1004002 MB1004003 MB1004004 MB1004005		Cow *(Bos primigenius)*	Hanmer Springs	NGS008	B4.8	GS454-01B
CMB05230 CMB05231 CMB05232 CMB05233 CMB05234	14.6	Duck (Anatidae)	Hagley Park, Christchurch	NGS010	B4.10	GS454-01B
CMB09197 CMB09198 CMB09199 CMB09200 CMB092001	16.07 25.58	Swan (*Cygnus)*	Bromley, Christchurch	NGS012	B4.12	GS454-01B
Cawthron 119	16.33	Human sewage	Northland	NGS013	B4.13	GS454-01B
CMB05123	21.9	Human sewage	Bromley	NGS014	B4.14	GA454-01B
CMB06668	22.2	Human sewage	Bromley	NGS015	B4.15	GS454-01B

Table 3.1 continued

ESR Sample ID	Previous qPCR results	Species	Location sample taken from	Study Sample ID	Barcode tag	Sequencing Run
CMB05201 CMB05202 CMB05203 CMB05204 CMB05205	17.65	Chicken *(Gallus gallus domesticus)*	Christchurch	NGS021	H2	GS454-02
CMB05207 CMB05208 CMB05209 CMB05210 CMB05211	11.54 11.91 11.52 12.19	Chicken *(Gallus gallus domesticus)*	Te Awamutu	NGS028	H16	GS454-02
CMB05214 CMB05215 CMB05216 CMB05217 CMB04118	16.94	Chicken *(Gallus gallus domesticus)*	Invercargill	NGS029	H17	GS454-02
CMB120280* CMB120281* CMB120283* CMB120284* CMB120287*	15.46 16.08 15.50 16.35 15.16	Sheep (*Ovis aries)*	Kaikoura	NGS031	H3	GS454-02
MB1104002 MB1104003 MB1104004 MB1104005 MB1104006		Cow *(Bos primigenius)*	Lincoln	NGS038	H4	GS454-02
Cawthron 51 Cawthron 52 Cawthron 53 Cawthron 64 Cawthron 67		Horse *(Equus ferus cabullus)*	Unknown	NGS046	H5	GS454-03
CMB07345 CMB07346 CMB07347 CMB07348 CMB07349	12.75 15.92 12.62 13.59	Pig *(Sus domesticus)*	Unknown	NGS048	H20	GS454-03
CMB05156 CMB05157 CMB05158 CMB05165 CMB05166		Dog (*Canis lupus familiaris)*	Unknown	NGS057	H19	GS454-02
CMB091305 CMB091306 CMB091307 CMB091308 CMB091309		Dog (*Canis lupus familiaris)*	Unknown	NGS059	H7	GS454-02
MB1204032 MB1204033 MB1204034 MB1204035 MB1204037	13.71 23.4 17.3	Dog (*Canis lupus familiaris)*	Unknown	NGS060	H21	GS454-02
CMB05031 CMB05032 CMB05033 CMB05034 CMB05035	19.1	Duck (Anatidae)	Hagley Park, Christchurch	NGS079	H8	GS454-03
CMB06216 CMB06217 CMB06218 CMB06219 CMB06220	25.7 22.23 21.55 25.6	Duck (Anatidae)	Hagley Park, Christchurch	NGS081	H6	GS454-02

Table 3.1 continued

ESR Sample ID	Previous qPCR results	Species	Location sample taken from	Study Sample ID	Barcode tag	Sequencing Run
CMB130002* CMB130003* CMB130005* CMB130006* CMB130008*		Duck (Anatidae)	The Groynes, Christchurch	NGS083	H24	GS454-02
CMB130009* CMB130010* CMB130011* CMB130012* CMB130013*		Duck (Anatidae)	The Groynes, Christchurch	NGS084	H25	GS454-02
MB1104020 MB1104021 MB1104022 MB1104023 MB1104024		Chicken *(Gallus gallus domesticus)*	Greymouth	NGS102	H14	GS454-02
CMB05195 CMB05196 CMB05197 CMB05198 CMB05199	16.67 16.14 14.88 36.41 16.67	Chicken *(Gallus gallus domesticus)*	Invercargill	NGS103	H15	GS454-02
CMB05069 CMB05070 CMB05071 CMB05072 CMB05073	18.49 16.19 18.57 17.28 18.04	Pukeko *(Porphyrio porphyria melanotus)*	Orana Park, Christchurch	NGS114	H9	GS454-03
CMB07300 CMB07301 CMB07302 CMB07303 CMB07304	11.73 18.36 14.84 12.33 12.44	Human (*Homo sapiens)*	Unknown	NGS119	H12	GS454-02
CMB09231 CMB09232 CMB09234 CMB09235 CMB09237	14.3 20.5	Human (*Homo sapiens)*	Unknown	NGS120	H13	GS454-02
CMB06589 CMB06591 CMB06593 CMB06600 CMB06602	16.7 17.9 17.6 17.3 17.0	Possum *(Trichosurus vulpecula)*	Lincoln	NGS136	H10	GS454-03
CMB120832 CMB120836 CMB120988 CMB120989	15.4 17.47 14.73 15.69	Alpaca *(Vicugna pacos)*	Unknown	NGS137	H11	GS454-03

Table 3.2: Water samples used in the combined sequencing study.

ESR Sample ID	Location sample taken from	Previous ESR contamination analysis outcome	Study Sample ID	Barcode tag	Sequencing Run
CMB120274	Auckland	Human	NGS016	B4.16	GS454-01B
CMB120322	Northland	Ruminant	NGS017	B4.17	GS454-01B
CMB120397	Southland	Ruminant	NGS018	B4.18 H7	GS454-01B GS454-03
CMB120346	Auckland	Human, Duck	NGS125	H26	GS454-03
CMB120351	Auckland	Ruminant	NGS126	H27	GS454-03
CMB120354	Auckland	Human, Dog	NGS127	H28	GS454-03
CMB120477	Auckland	Human	NGS128	H29	GS454-03
CMB120701	Canterbury	Ruminant	NGS129	H30	GS454-03
CMB120750	Northland	Ruminant	NGS130	H31	GS454-03
CMB120751	Northland	Human	NGS131	H12 H32	GS454-03

Faecal and water samples were prepared for amplicon sequencing as outlined in Chapter 2, Sections 2.3.1 and 2.3.2, with the following minor changes to the protocol.

3.3.1.1 Oligonucleotide primer design

The hypervariable region V1-V3 target was kept, using the Bac8F (5'-AGAGTTTGATCCTGGCTCAG-3') and Univ529R (5'-ACCGCGGCKGCTGGC-3') universal eubacterial primer set (Fierer *et al.*, 2007). The sample specific barcodes were modified to be 8 nt in length, allowing for a greater confidence in correctly identifying sequence sources and to include the error-correcting ability Hamming-based barcodes provide. 30 barcode sequences were selected from the H4(8,4) coding provided by Bystrykh (2012) and were included at the 5' end of the universal primers (Table 3.3).

Table 3.3: PCR primers used for samples sequenced in GS454-02 and GS454-03. Barcode sequences are highlighted in bold.

Primer	Nucleotide sequence	T_m (°C)
Bac8F H2	**TCACAGCA**AGAGTTTGATCCTGGCTCAG	56
Univ529R H2	**TCACAGCA**ACCGCGGCKGCTGGC	60
Bac8F H3	**GTAGCACT**AGAGTTTGATCCTGGCTCAG	56
Univ529R H3	**GTAGCACT**ACCGCGGCKGCTGGC	60
Bac8F H4	**ATAGCGTC**AGAGTTTGATCCTGGCTCAG	56
Univ529R H4	**ATAGCGTC**ACCGCGGCKGCTGGC	60
Bac8F H5	**CTAGCTGA**AGAGTTTGATCCTGGCTCAG	56
Univ529R H5	**CTAGCTGA**ACCGCGGCKGCTGGC	60
Bac8F H6	**CTACGACA**AGAGTTTGATCCTGGCTCAG	56
Univ529R H6	**CTACGACA**ACCGCGGCKGCTGGC	60

Table 3.3 continued

Primer	Nucleotide sequence	T_m (°C)
Bac8F H7	**GTACGCAT**AGAGTTTGATCCTGGCTCAG	56
Univ529R H7	**GTACGCAT**ACCGCGGCKGCTGGC	60
Bac8F H8	**ACATGCGT**AGAGTTTGATCCTGGCTCAG	56
Univ529R H8	**ACATGCGT**ACCGCGGCKGCTGGC	60
Bac8F H9	**GCATGTAC**AGAGTTTGATCCTGGCTCAG	56
Univ529 H9	**GCATGTAC**ACCGCGGCKGCTGGC	60
Bac8F H10	**ATACGTGC**AGAGTTTGATCCTGGCTCAG	56
Univ529 H10	**ATACGTGC**ACCGCGGCKGCTGGC	60
Bac8F H11	**GCAGTATC**AGAGTTTGATCCTGGCTCAG	56
Univ529R H11	**GCAGTATC**ACCGCGGCKGCTGGC	60
Bac8F H12	**TCAGTCGA**AGAGTTTGATCCTGGCTCAG	56
Univ529R H12	**TCAGTCGA**ACCGCGGCKGCTGGC	60
Bac8F H13	**ACAGTGCT**AGAGTTTGATCCTGGCTCAG	56
Univ529R H13	**ACAGTGCT**ACCGCGGCKGCTGGC	60
Bac8F H14	**TGCTACAG**AGAGTTTGATCCTGGCTCAG	56
Univ529 H14	**TGCTACAG**ACCGCGGCKGCTGGC	60
Bac8F H15	**AGCTAGTC**AGAGTTTGATCCTGGCTCAG	56
Univ529 H15	**AGCTAGTC**ACCGCGGCKGCTGGC	60
Bac8F H16	**CGCTATGA**AGAGTTTGATCCTGGCTCAG	56
Univ529R H16	**CGCTATGA**ACCGCGGCKGCTGGC	60
Bac8F H17	**GACACTAC**AGAGTTTGATCCTGGCTCAG	56
Univ529R H17	**GACACTAC**ACCGCGGCKGCTGGC	60
Bac8F H19	**GACTGATC**AGAGTTTGATCCTGGCTCAG	56
Univ529R H19	**GACTGATC**ACCGCGGCKGCTGGC	60
Bac8F H20	**TACTGCGA**AGAGTTTGATCCTGGCTCAG	56
Univ529R H20	**TACTGCGA**ACCGCGGCKGCTGGC	60
Bac8F H21	**CACTGTAG**AGAGTTTGATCCTGGCTCAG	56
Univ529R H21	**CACTGTAG**ACCGCGGCKGCTGGC	60
Bac8F H22	**TGCATACG**AGAGTTTGATCCTGGCTCAG	56
Univ529R H22	**TGCATACG**ACCGCGGCKGCTGGC	60
Bac8F H23	**CACGTATG**AGAGTTTGATCCTGGCTCAG	56
Univ529R H23	**CACGTATG**ACCGCGGCKGCTGGC	60
Bac8F H24	**AGCATCAC**AGAGTTTGATCCTGGCTCAG	56
Univ529 H24	**AGCATCAC**ACCGCGGCKGCTGGC	60
Bac8F H25	**TACGTGCA**AGAGTTTGATCCTGGCTCAG	56
Univ529R H25	**TACGTGCA**ACCGCGGCKGCTGGC	60
Bac8F H26	**CGCATGTA**AGAGTTTGATCCTGGCTCAG	56
Univ529R H26	**CGCATGTA**ACCGCGGCKGCTGGC	60
Bac8F H27	**GCGTACAT**AGAGTTTGATCCTGGCTCAG	56
Univ529R H27	**GCGTACAT**ACCGCGGCKGCTGGC	60
Bac8F H28	**ATGCACGT**AGAGTTTGATCCTGGCTCAG	56
Univ529R H28	**ATGCACGT**ACCGCGGCKGCTGGC	60
Bac8F H29	**TCGTAGTG**AGAGTTTGATCCTGGCTCAG	56
Univ529R H29	**TCGTAGTG**ACCGCGGCKGCTGGC	60
Bac8F H30	**GTGCATAC**AGAGTTTGATCCTGGCTCAG	56
Univ529 H30	**GTGCATAC**ACCGCGGCKGCTGGC	60
Bac8F H31	**ACGTATGC**AGAGTTTGATCCTGGCTCAG	56
Univ529R H31	**ACGTATGC**ACCGCGGCKGCTGGC	60
Bac8F H32	**GTGACATC**AGAGTTTGATCCTGGCTCAG	56
Univ529R H32	**GTGACATC**ACCGCGGCKGCTGGC	60

3.3.1.2 PCR amplification of DNA targets

A recent study assessed amplification efficiency of a range of polymerase enzymes for use in NGS applications, concluding that Kapa HiFi polymerase (Kapa BioSystems, USA) was the best enzyme for NGS library preparations (Quail *et al.*, 2012a). This enzyme was trialled with a number of faecal and water samples and compared against Platinum Taq High Fidelity Polymerase (Invitrogen, USA). Using UV visualisation of agarose gels, Kapa HiFi provided stronger amplification for all samples trialled, with the exception of bird samples, which had slightly weaker amplification compared to the Invitrogen Taq (data not shown). Kapa HiFi enzymes were therefore used for the GS454-02 and GS454-03 amplicon library preparation, following the reaction mix listed in Table 3.4.

Table 3.4: PCR reaction mix for samples sequenced in GS454-02 and GS454-03.

Reagent	Concentration per reaction tube	Volume per reaction tube (µl)
5x Buffer (includes 2 mM Mg^{2+})	1x	5
Kapa dNTPs (10 mM each)	0.3 mM	0.75
Kapa HiFi polymerase (1 unit/µl)	0.5 units	0.5
Primers (10 pmol/µl)	0.3 µM each	0.75 (of each)
DNA		1
dH_2O		16.25
Total		25

PCR amplification was initiated with a denaturation step at 95°C for 2 min, followed by a three stage programme of 25 repeated cycles. Each amplification cycle consisted of a denaturation step (98°C for 20 s), an annealing step (68°C for 15 s) and an extension step (72°C for 15 s). A second extension step of 72°C for 3 min followed these 25 cycles. A final step of 20°C was included to keep reactions at room temperature until processing. Due to the small amplification volume, amplicons were prepared in triplicate, and pooled prior to purification.

3.3.1.3 Pooling of amplicons

Amplicons were prepared for two sequencing samples, each containing 16 samples (Tables 3.1 and 3.2). Individual samples were diluted in molecular biology grade water to the lowest concentration of the samples to be pooled, 10 ng/µl and 14 ng/µl for

GS454-02 and GS454-03 sequencing samples, respectively. 5 µl of each diluted sample were pooled together to provide the final sample to be sent for sequencing.

3.3.2 Next generation sequencing

GS454-01B was sequenced as outlined in Chapter 2 (Section 2.3.3). For GS454-02 and GS454-03, ligation of 454 sequencing adaptors and sequencing of the pooled PCR products was provided by Macrogen Inc. (Korea), using a 1/8 region plate of a Roche 454 GS FLX platform for each sequencing sample.

3.3.3 Data analysis

The raw data provided for GS454-01B (NZGL, New Zealand) was initially imported into Geneious (Biomatters, New Zealand) to remove the 454 sequencing adaptors, and was exported as a FASTA file with a matching quality file. The raw data provided for GS454-02 and GS454-03 (Macrogen Inc., Korea) already had the sequencing adaptors removed and provided both the FASTA and quality files. These files were used to initiate the QIIME pipeline, as outlined in Chapter 2, Section 2.3.4.3. New mapping files were created for the two new sequencing runs, and are included in Appendix II. The data from each sequencing run was filtered and sorted based on barcode, and five additional samples included that were not part of this analysis were removed. All the sequence files were then combined to complete the remainder of the analysis as one data set.

The "pick_otus_through_otu_table.py" workflow script was replaced with the independent scripts to enable the addition of a chimera checking script, using ChimeraSlayer (Haas *et al.*, 2011). This removes sequences which can be assigned to multiple taxonomies, generally due to having multiple parent sequences during the amplification steps, which can have a significant impact on diversity measures (Kunin *et al.*, 2010; Schloss *et al.*, 2011).

The remainder of the analyses were completed as outlined in Section 2.3.4.3.

3.4 Results

3.4.1 DNA sequencing

The GS454-01B sample contained 20 samples at equimolar ratios (Figure 2.5). A raw data sff file was provided by Auckland University, containing 212,106 sequences. GS454-02 and GS454-03 sequencing samples each contained 16 samples at equimolar ratios (Figures 3.1 and 3.2). Raw data from Macrogen Inc. included a FASTA file and a quality file, with all the sequencing adaptors removed, containing 168,400 and 141,559 sequences for GS454-02 and GS454-03, respectively, resulting in a raw dataset of 522,065 sequences, divided across 52 samples.

3.4.2 Data analysis

3.4.2.1 Data filtering and OTU selection

The data from the filtering steps provided by the "split_libraries.py" script for each of the sequencing samples are summarised in Tables 3.5 and 3.6. After the filtering steps, the sequences from all three sequencing samples were combined into one file for further analysis, consisting of 389,963 sequences. Sequences from five samples included in the sequencing were removed prior to OTU selection, including NGS003.B4.3 and NGS005.B4.5, which were amplified using a different amplification protocol, and three aquifer samples, which were not part of this analysis.

After removing chimeric sequences, 39,112 OTU sequences were identified, and classified using the RDP Classifier, with 342,616 individual observations (Table 3.6). The minimum number of OTUs assigned to a sample was 92; the maximum 16,318, with a mean of 7,290.

Figure 3.1: Agarose gel of final amplicon samples for GS454-02 sequencing. All samples are at a concentration of 10 ng/µl. Lane one contains 1 kb plus DNA ladder (Invitrogen).

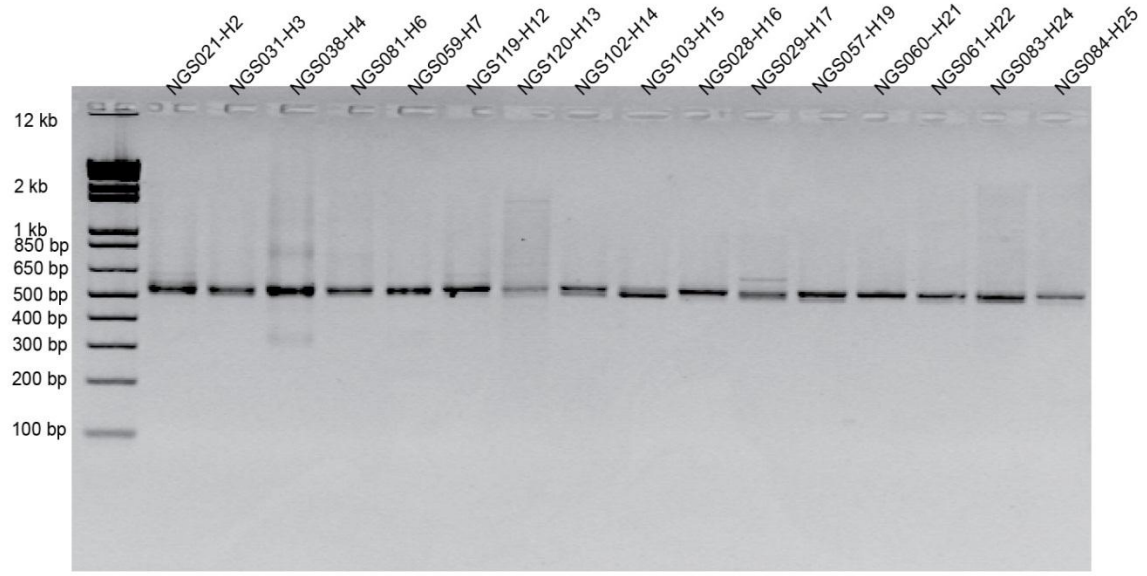


Figure 3.2: Agarose gel of final amplicon samples for GS454-03 sequencing. All samples are at a concentration of 14 ng/µl. Lane one contains 1 kb plus DNA ladder (Invitrogen).

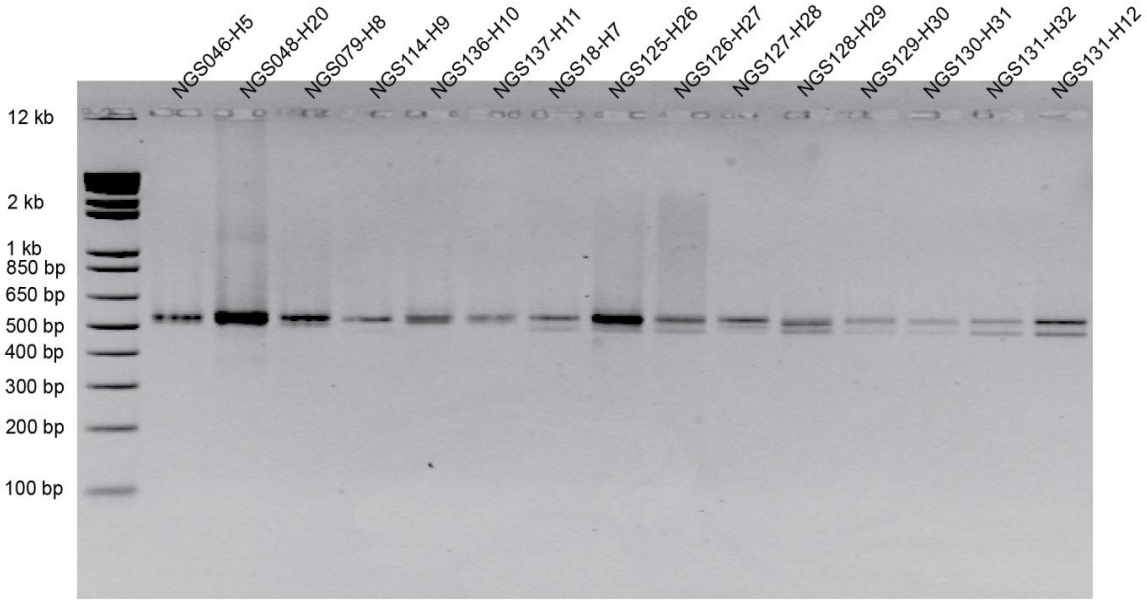


Table 3.5: Initial processing and filtering of raw data.

	GS454-01B		GS454-02		GS454-03	
	Forward map	**Reverse map**	**Forward map**	**Reverse map**	**Forward map**	**Reverse map**
Raw input sequences	212106		168400		141559	
Failed size check	26212 (12.4%)		13940 (8.3%)		9420 (6.7%)	
Failed ambiguous bases	0 (0%)		8 (0.0%)		42 (0.0%)	
Failed mean quality score	14909 (7.0%)		3633 (2.2%)		1897 (1.3%)	
Failed homopolymers	1624 (0.8%)		2154 (1.3%)		2383 (1.7%)	
No primer match	98825 (46.6%)	85416 (40.3%)	74029 (44.0%)	94693 (56.2%)	69822 (49.3%)	78065 (55.1%)
Barcodes not in mapping file	n/a	n/a	110 (0.1%)	101 (0.1%)	466 (0.3%)	99 (0.1%)
Total sequences written to file	70481 (33.2%)	83902 (39.6%)	74526 (44.3%)	53869 (32.0%)	57529 (40.6%)	49653 (35.1%)
Total concatenated sequences	154386 (72.8%)		128395 (76.2%)		107182 (75.7%)	
Minimum no. sequences/sample	47	51	837	523	120	183
Maximum no. sequences/sample	8378	9525	9348	7536	8334	7409

Table 3.6: Numbers of sequences and OTUs assigned to each barcoded sample.

	Forward map	Reverse map	Combined	Total OTUs	Faecal OTUs
NGS001-B4.1	6491	8565	15056	14799	12461
NGS002-B4.2	2544	4063	6607	6445	5596
NGS003-B4.9	4349	5121	9470	9296	7769
NGS004-B4.4	4449	5815	10260	10071	8681
NGS005-B4.11	2811	3006	5817	5729	4892
NGS006-B4.6	475	599	1074	1060	913
NGS007-B4.7	2583	3454	6037	5865	5130
NGS008-B4.8	2955	3334	6289	6108	5365
NGS010-B4.10	5624	5852	11476	10303	9632
NGS012-B4.12	3898	5394	9292	7801	7464
NGS013-B4.13	6456	7183	13639	12663	7495
NGS014-B4.14	8378	9001	17379	16318	12526
NGS015-B4.15	6565	7640	14205	12509	10212
NGS016-B4.16	461	556	1017	869	455
NGS017-B4.17	534	669	1203	1117	339
NGS018-B4.18	47	51	98	92	44
NGS018.H7	5084	4725	9809	9556	5633
NGS021.H2	1458	1041	2499	2368	1726
NGS028.H16	3922	2212	6134	5815	4662
NGS029.H17	3610	3684	7294	6904	4899
NGS031.H3	8208	5817	14025	13820	12057
NGS038.H4	5545	3138	8683	8592	7267
NGS046.H5	2934	2506	5440	5365	4472
NGS048.H20	4738	3883	8621	8227	7400
NGS057.H19	9348	7061	16409	15921	14896
NGS059.H7	8658	5158	13816	13773	13228
NGS060.H21	4274	3469	7743	7695	7377
NGS061.H22	989	523	1512	1502	1403
NGS079.H8	6975	5528	12503	11793	10825
NGS081.H6	3069	2285	5354	4926	4492
NGS083.H24	8390	5462	13852	12471	11261
NGS084.H25	922	635	1557	1416	1346
NGS102.H14	1910	1376	3286	3223	2243
NGS103.H15	8572	7536	16108	15423	11165
NGS114.H9	4247	3578	7825	7786	6470
NGS119.H12	837	681	1518	1442	1335
NGS120.H13	4814	3791	8605	8180	7621
NGS125.H26	5045	4241	9286	8876	4539
NGS126.H27	3913	3599	7512	7225	2275
NGS127.H28	3876	3532	7408	7229	3984
NGS128.H29	120	183	303	300	215
NGS129.H30	2233	1538	3771	3590	1723
NGS130.H31	945	473	1418	1385	879
NGS131.H12	512	478	990	4091	2612
NGS131.H32	2033	2134	4167	955	569
NGS136.H10	3496	2866	6362	6234	5579
NGS137.H11	8334	7409	15743	15488	13506
Sequences included but not analysed (5 samples)	11160	12175	23335	-	-

3.4.2.2 Taxonomy classifications

342,616 individual OTUs were classified using the RDP Classifier through the QIIME pipeline. 6% were classified as not being bacteria, and were removed from further analyses.

3.4.2.2.1 Phyla level classifications

Classifications at the phyla level are shown for each faecal library and water sample in Figures 3.3 and 3.4, respectively.

Among faecal samples *Bacteroidetes* and *Firmicutes* were generally the most common phyla. *Bacteroidetes* ranged from 25% in chicken faeces up to 54% in duck faeces, while *Firmicutes* ranged from 30% in dog faeces up to 54% in pukeko faeces. The exception was the swan faeces sample, with just 18% classified as *Bacteroidetes* and 4.5% *Firmicutes*. Instead swan faeces had a high proportion of *Fusobacteria* (37%), and *Proteobacteria* (36%). *Fusobacteria* were also prominent in dog faeces (15%), while chicken, cow, duck, human pig and sheep samples all had less than 2%. *Tenericutes* were also present in all samples, with pukeko containing the highest percentage (6.7%). Municipal sewage contained 61% *Proteobacteria*, 23% *Bacteroidetes*, 8% *Firmicutes* and 3% *Fusobacteria*. 27 different phyla identified are considered to be candidate divisions (McDonald *et al.* 2012) so were combined together (Candidate divisions), with the majority of these only found in the water samples. An additional six phyla that were not identified in the previous dataset (Chapter 2) were also combined together (labelled 'Other phyla'). Individually, none of these make up more than 1% of the total number of sequences.

In the water samples, *Bacteroidetes* and *Proteobacteria* were the most common phyla. *Bacteroidetes* ranged from 14% to 53%, while *Proteobacteria* ranged from 25% up to 62%. *Actinobacteria, Cyanobacteria* and *Firmicutes* were also found in all samples, although generally at much lower numbers, from below 1% up to 16%. A much broader range of Candidate divisions were also seen throughout the samples, accounting for up to 16%.

3.4.2.2.2 Genus level classifications

A 5% threshold was set for evaluating genus level classifications for the identification of potential MST markers, or an OTU classification had to be present in only one faecal source. All potential markers were found in the four dominant phyla, with only one

genus below 5% that was only present in one faecal source (Table 3.7). Five genera from *Bacteroidetes* were found to be dominant, although all were found within multiple species sources. Nine dominant genera were found within *Firmicutes*, with seven of these found strongly in only one source species. Two different genera within *Fusobacteria* account for most of this phyla's representation in the swan and dog sources; five different genera were found from the *Proteobacteria* phyla, with all of these only dominant in one source, three of which were found in sewage.

3.4.2.2.3 Water sample analysis

Table 3.8 provides the percentage of OTUs per water sample for the identified potential MST markers. Based on the analysis of the genera and which faecal source each were found in, a number of the water samples can be loosely classified as to their faecal contamination (Table 3.9). Using a 1% detection threshold, contamination sources can be attributed to six of the ten different water samples, classified as being predominantly contaminated by either human or ruminant sources based on the presence of identified source-specific genera.

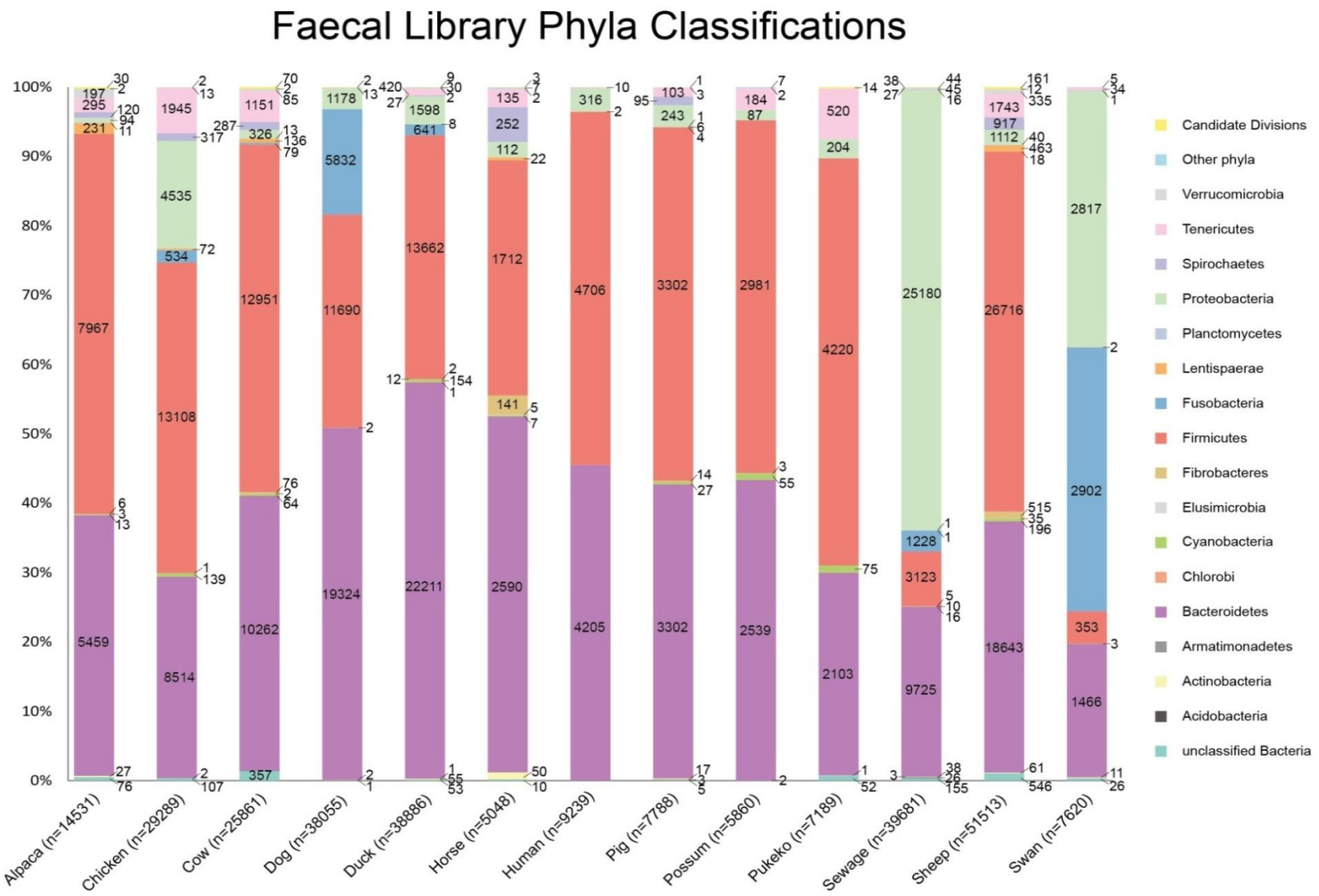


Figure 3.3: Phyla taxonomy level classifications for faecal source library samples. Figures in each bar represent the number of OTUs assigned to the corresponding phylum.

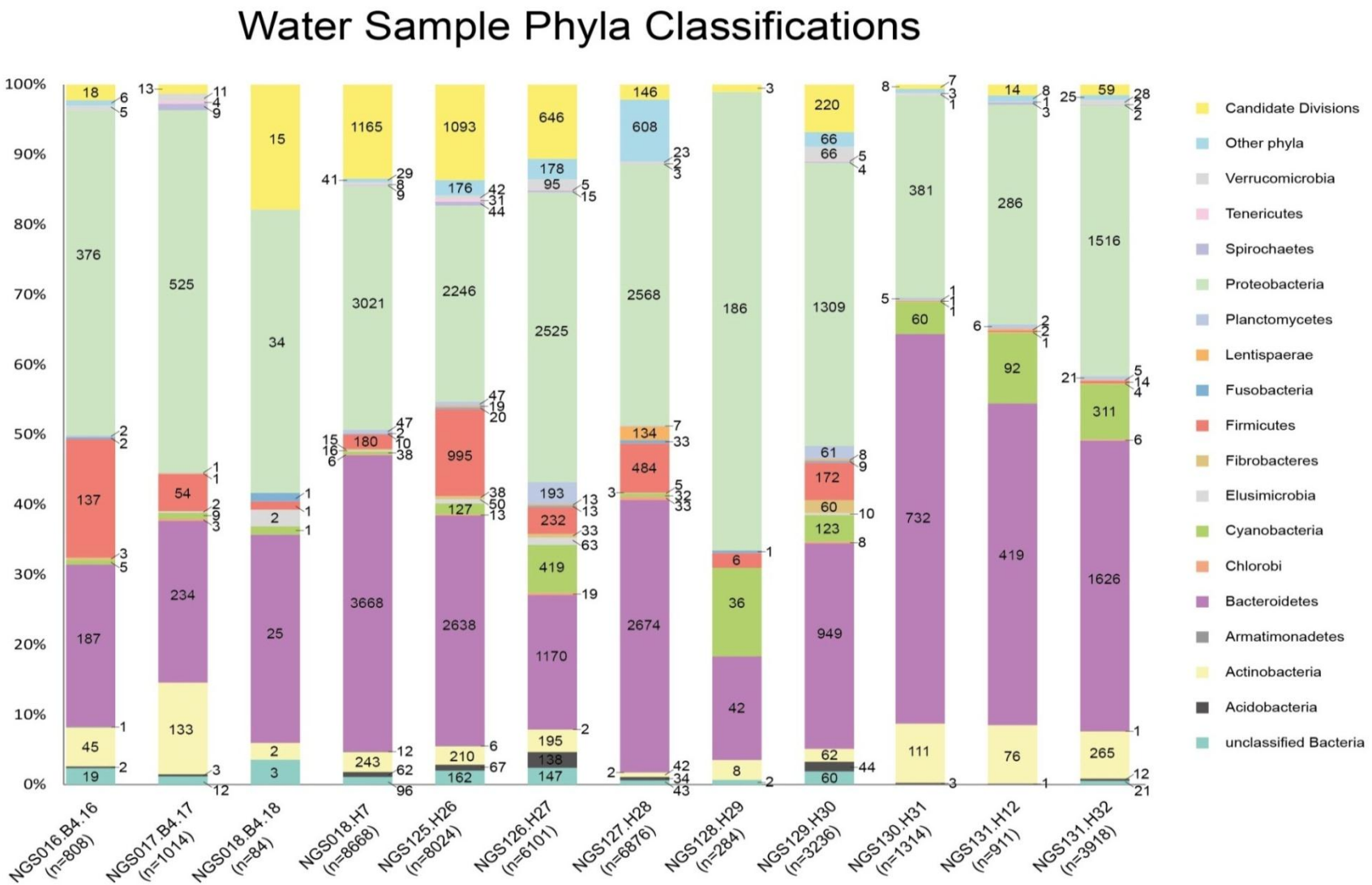


Figure 3.4: Phyla taxonomy level classifications for water samples. Figures in each bar represent the number of OTUs assigned to the corresponding phylum**.**

Table 3.7: Potential genus-level markers from faecal sources. Percentages greater than 5%, or where only found in one source library, are highlighted in bold.

		Ruminant animals			Non-ruminant animals				Human		Birds			
	Predominant source(s)	**Alpaca**	**Cow**	**Sheep**	**Dog**	**Horse**	**Pig**	**Possum**	**Human**	**Sewage**	**Chicken**	**Duck**	**Pukeko**	**Swan**
Bacteroidetes; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacteroides	-	1.4%	2.6%	4.0%	**18.5%**	0.8%	1.3%	**13.0%**	**35.6%**	**6.0%**	**18.0%**	**24.0%**	**9.5%**	**18.0%**
Firmicutes; Clostridia; Clostridiales; Ruminococcaceae; Oscillospira	-	2.3%	1.8%	2.6%	0.07%	1.1%	2.4%	**6.3%**	4.9%	0.4%	**6.1%**	4.8%	2.9%	0.3%
Firmicutes; Clostridia; Clostridiales; Ruminococcaceae; Ruminococcus	-	**21.3%**	**21.9%**	**18.9%**	0.1%	4.0%	2.1%	**7.0%**	1.9%	0.3%	1.7%	1.1%	**6.4%**	0.5%
Bacteroidetes; Bacteroidia; Bacteroidales; Porphyromonadaceae; Parabacteroides	Human	0.5%	0.08%	0.02%	0.6%		2.1%	3.8%	**7.3%**	1.3%	0.3%	1.4%	-	-
Firmicutes; Clostridia; Clostridiales; Lachnospiraceae; Blautia	Human	-	-	-	1.7%	0.02%	0.2%	-	**6.1%**	0.3%	0.2%	-	-	-
Firmicutes; Clostridia; Clostridiales; Lachnospiraceae; Roseburia	Human	-	-	-	0.02%	0.3%	0.1%	-	**7.9%**	0.4%	0.03%	-	-	-
Firmicutes; Clostridia; Clostridiales; Ruminococcaceae; Faecalibacterium	Human (Duck, dog)	-	-	-	2.3%	-	0.09%	1.0%	**7.7%**	0.8%	1.0%	3.9%	-	0.04%
Proteobacteria; Betaproteobacteria; Burkholderiales; Alcaligenaceae; Sutterella	Chicken (Human, dog)	0.05%	0.2%	0.08%	2.8%	-	0.2%	0.2%	2.8%	0.4%	**5.8%**	1.0%	-	0.2%
Proteobacteria; Betaproteobacteria; Rhodocyclales; Rhodocyclaceae; Zoogloea	Sewage	-	-	-	-	-	-	-	-	**1.4%**	-	-	-	-
Proteobacteria; Betaproteobacteria; Burkholderiales; Comamonadaceae; Acidovorax	Sewage	-	-	-	-	-	-	-	-	**6.9%**	-	-	-	-
Proteobacteria; Epsilonproteobacteria; Campylobacterales; Campylobacteraceae; Arcobacter	Sewage	-	-	-	-	-	0.6%	-	-	**30.5%**	-	-	-	0.2%
Bacteroidetes; Flavobacteria; Flavobacteriales; Flavobacteriaceae; Flavobacterium	Horse, Sewage	-	-	0.04%	-	**5.4%**	0.4%	-	-	**5.2%**	-	-	-	0.3%
Firmicutes; Bacilli; Lactobacillales; Lactobacillaceae; Lactobacillus	Pig (Chicken)	-	-	-	0.1%	0.06%	**20.7%**	-	0.6%	0.03%	2.8%	0.2%	-	0.01%
Bacteroidetes; Bacteroidia; Bacteroidales; Bacteroidaceae; 5-7N15	Ruminant	**6.3%**	**5.0%**	3.9%	-	-	-	-	-	0.01%	-	0.01%	-	0.01%
Firmicutes; Clostridia; Clostridiales; Veillonellaceae; Megamonas	Dog (Chicken, duck)	-	-	-	**11.0%**	-	0.05%	-	0.07%	0.06%	2.2%	2.1%	-	0.03%
Firmicutes; Clostridia; Clostridiales; Clostridiaceae; Sarcina	Dog	-	-	-	**5.6%**	0.04%	-	-	-	-	-	-	-	-

Table 3.7 continued

	Predominant source(s)	Ruminant animals			Non-ruminant animals				Human		Birds			
		Alpaca	Cow	Sheep	Dog	Horse	Pig	Possum	Human	Sewage	Chicken	Duck	Pukeko	Swan
Fusobacteria; Fusobacteria; Fusobacteriales; Fusobacteriaceae; J2-29	Dog	-	-	-	**9.4%**	-	-	-	-	-	-	-	-	-
Bacteroidetes; Bacteroidia; Bacteroidales; Prevotellaceae; Prevotella	Dog, Duck, Pig	3.3%	0.2%	0.3%	**25.7%**	1.7%	**13.0%**	0.03%	2.3%	3.6%	0.2%	**14.3%**	-	0.01%
Firmicutes; Bacilli; Bacillales; Planococcaceae; Solibacillus	Horse	1.3%	0.03%	0.8%	-	**7.1%**	0.01%	-	-	-	-	-	-	-
Fusobacteria; Fusobacteria; Fusobacteriales; Fusobacteriaceae; Cetobacterium	Swan	-	0.3%	0.03%	-	-	-	-	-	0.03%	-	-	-	**36.3%**
Proteobacteria; Gammaproteobacteria; Alteromonadales; Shewanellaceae; Shewanella	Swan	-	1.1%	0.02%	-	-	-	-	-	0.1%	-	-	-	**19.4%**

Table 3.8: Presence of potential genus-level markers in water samples. Percentages greater than 1% are highlighted in bold.

	Predominant source(s)	Human	Human	Human	Human	Human, Duck	Human, Dog	Ruminant	Ruminant	Ruminant	Ruminant	Ruminant	Ruminant
	Sample	NGS016	NGS128	NGS131 H12	NGS131 H32	NGS125	NGS127	NGS017	NGS018 B4.18	NGS018 H7	NGS126	NGS129	NGS130
Bacteroidetes; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacteroides	-	**10.4%**	0.7%	0.1%	0.1%	**10.8%**	**7.4%**	0.4%	-	0.4%	**1.1%**	0.9%	-
Firmicutes; Clostridia; Clostridiales; Ruminococcaceae; Oscillospira	-	0.3%	-	-	-	0.3%	0.03%	0.2%	-	0.06%	0.07%	0.1%	-
Firmicutes; Clostridia; Clostridiales; Ruminococcaceae; Ruminococcus	-	0.5%	-	0.1%	0.05%	0.3%	0.06%	**1.5%**	-	0.4%	0.2%	**1.7%**	-
Bacteroidetes; Bacteroidia; Bacteroidales; Porphyromonadaceae; Parabacteroides	Human	0.5%	0.7%	-	0.05%	**1.3%**	**5.9%**	-	-	0.08%	0.08%	0.06%	0.07%
Firmicutes; Clostridia; Clostridiales; Lachnospiraceae; Blautia	Human	**1.4%**	-	-	-	0.4%	0.1%	-	-	-	-	-	-
Firmicutes; Clostridia; Clostridiales; Lachnospiraceae; Roseburia	Human	-	**1.0%**	-	-	**1.3%**	-	-	-	0.01%	-	-	-
Firmicutes; Clostridia; Clostridiales; Ruminococcaceae; Faecalibacterium	Human (Duck, dog)	**3.2%**	-	-	-	0.7%	0.1%	-	-	-	0.04%	-	-
Proteobacteria; Betaproteobacteria; Burkholderiales; Alcaligenaceae; Sutterella	Chicken (Human, dog)	-	0.3%	-	-	0.1%	0.07%	-	-	-	-	0.03%	-

Table 3.8: continued

	Predominant source(s)	Human	Human	Human	Human	Human, Duck	Human, Dog	Ruminant	Ruminant	Ruminant	Ruminant	Ruminant	Ruminant
	Sample	**NGS016**	**NGS128**	**NGS131 H12**	**NGS131 H32**	**NGS125**	**NGS127**	**NGS017**	**NGS018 B4.18**	**NGS018 H7**	**NGS126**	**NGS129**	**NGS130**
Proteobacteria; Betaproteobacteria; Rhodocyclales; Rhodocyclaceae; Zoogloea	Sewage	0.5%	0.7%	-	0.02%	0.4%	**2.3%**	-	-	-	-	0.03%	-
Proteobacteria; Betaproteobacteria; Burkholderiales; Comamonadaceae; Acidovorax	Sewage	**1.3%**	0.3%	0.2%	0.2%	0.3%	0.9%	-	-	0.2%	0.01%	0.03%	-
Proteobacteria; Epsilonproteobacteria; Campylobacterales; Campylobacteraceae; Arcobacter	Sewage	**3.6%**	**42.7%**	0.9%	0.3%	0.7%	3.0%	0.09%	-	0.03%	0.4%	0.1%	0.07%
Bacteroidetes; Flavobacteria; Flavobacteriales; Flavobacteriaceae; Flavobacterium	Horse, Sewage	**1.6%**	**1.3%**	**1.7%**	**2.9%**	**3.7%**	**5.5%**	**8.7%**	**5.4%**	**10.6%**	**3.9%**	**5.5%**	**10.3%**
Firmicutes; Bacilli; Lactobacillales; Lactobacillaceae; Lactobacillus	Pig (Chicken)	-	-	-	-	0.05%	-	-	-	-	0.1%	-	-
Bacteroidetes; Bacteroidia; Bacteroidales; Bacteroidaceae; 5-7N15	Ruminant	0.1%	-	-	-	-	-	0.2%	**1.1%**	0.2%	0.03%	0.9%	-
Firmicutes; Clostridia; Clostridiales; Veillonellaceae; Megamonas	Dog (Chicken, duck)	-	-	-	-	0.1%	0.2%	-	-	-	0.2%	-	-
Firmicutes; Clostridia; Clostridiales; Clostridiaceae; Sarcina	Dog	-	-	-	-	-	-	-	-	-	-	-	-
Fusobacteria; Fusobacteria; Fusobacteriales; Fusobacteriaceae; J2-29	Dog	-	-	-	-	-	0.01%	-	-	-	-	-	-
Bacteroidetes; Bacteroidia; Bacteroidales; Prevotellaceae; Prevotella	Dog, Duck, Pig	0.6%	2.3%	0.4%	0.02%	0.8%	0.6%	-	-	0.4%	0.5%	0.1%	-
Firmicutes; Bacilli; Bacillales; Planococcaceae; Solibacillus	Horse	-	-	-	-	-	-	0.09%	-	0.01%	-	0.08%	-
Fusobacteria; Fusobacteria; Fusobacteriales; Fusobacteriaceae; Cetobacterium	Swan	-	-		-	0.05%	-	-	-	-	0.07%	-	-
Proteobacteria; Gammaproteobacteria; Alteromonadales; Shewanellaceae; Shewanella	Swan	0.1%	-	-	-	-	-	-	-	-	-	-	-

Table 3.9: Potential contamination sources of water samples using genus level markers. Genera highlighted in bold are the suggested contamination source, with genera in brackets found in levels below 1%.

Water sample	Potential human faecal markers present	Potential sewage markers	Potential ruminant markers present	Potential dog markers present	Significant genera not assigned	Commercial analysis results[1]
NGS016	***Blautia, Faecalibacterium, (Parabacteroides)***	***Acidovorax, Arcobacter, (Zoogloea)***	*(Ruminococcus)*	-	*Bacteroides, Flavobacterium*	Human
NGS128	*Roseburia, (Parabacteroides)*	***Arcobacter, (Acidovorax), (Zoogloea)***	-	-	*Prevotella, Flavobacterium*	Human
NGS131	*(Parabacteroides)*	*(Acidovorax), (Zoogloea), (Arcobacter)*	*(Ruminococcus)*		*Flavobacterium*	Human
NGS125	***Parabacteroides, Roseburia, (Blautia), (Faecalibacterium)***	*(Acidovorax), (Zoogloea), (Arcobacter)*	*(Ruminococcus)*	*(Megamonas)*	*Bacteroides, Flavobacterium*	Human, duck
NGS127	***Parabacteroides, (Blautia), (Faecalibacterium)***	***Zoogloea, Arcobacter, (Acidovorax)***	*(Ruminococcus)*	*(Megamonas), (J2-29)*	*Bacteroides, Flavobacterium*	Human, dog
NGS017	-	*(Arcobacter)*	***Ruminococcus (5-7N15)***	-	*Flavobacterium*	Ruminant
NGS018	*(Parabacteroides)*	*(Arcobacter)*	***5-7N15, Ruminococcus***	-	*Flavobacterium*	Ruminant
NGS126	*(Parabacteroides), (Faecalibacterium)*	*(Acidovorax), (Arcobacter)*	*(5-7N15) (Ruminococcus)*	*(Megamonas)*	*Bacteroides, Flavobacterium*	Ruminant
NGS129	*(Parabacteroides)*	*(Acidovorax), (Zoogloea), (Arcobacter)*	*Ruminococcus, (5-7N15)*	-	*Flavobacterium*	Ruminant
NGS130	*(Parabacteroides)*	*(Arcobacter)*	-	-	*Flavobacterium*	Ruminant

3.4.2.3 Microbial community diversity

Microbial community diversity was measured through both α- and β-diversity metrics. α-diversity examined the microbial diversity within a sample, while β-diversity assessed the diversity between a collection of samples.

3.4.2.3.1 α-diversity

A range of metrics can be used to calculate the microbial diversity within a community, which reflects the diversity based on the abundance of taxa within each community. Four metrics were included in this study, chosen to include a range of measurement

[1] Sterol and/or PCR MST analysis

indices, including species-based qualitative indices (Chao1 and Observed species), a qualitative divergence-based index (Phylogenetic distance) and a quantitative species-based index (Shannon), which combined both species richness and evenness (Lozupone and Knight, 2008). QIIME creates plots of alpha diversity verses simulated sequencing effort, known as rarefaction plots. Rarefaction plots were generated for the four α-diversity metrics, and grouped according to source species (Figure 3.5), with water samples analysed individually (Figure 3.6). For sources with multiple samples, the rarefaction curve plots the average diversity, with error bars displaying the variation between the different samples. A sampling depth of 500 was used, as this provided a good sampling depth for the majority of the samples. Two water samples (NGS018-B4.18 and NGS128.H29) did not reach this diversity level, so consequently have a shorter rarefaction curve.

Rarefaction compares observed richness among samples which have been unequally sampled (Hughes *et al.*, 2001), as is the case with NGS which results in different numbers of sequences for each sample. The curves contain information about how well the communities have been sampled, with diversity considered to be well sampled once the curve plateaus and additional sampling does not contribute further to the total number of OTUs (Hughes *et al.*, 2001; Wooley *et al.*, 2010). The overall trend for all four metrics is the same, showing that the sheep, cow, alpaca and horse have the highest diversities, and in three of the four metrics do not appear to be reaching a plateau by 500 sequences. The other species all have a much lower diversity, and each curve appears to begin to plateau out to some extent by 500 sequences. The dog, swan and human samples consistently show the lowest diversity. The Phylogenetic distance index, which was the only divergence-based metric used, shows the same overall trends as the three species-based indices.

Individual rarefaction curves were generated for each water sample (Figure 3.6). Overall, the four different metrics show similar trends to those shown for the faecal samples, although the curves do not show the same extent of levelling off as for many of the faecal sources, suggesting there is more diversity that has not been sampled. NGS125, NGS126 and NGS129 show the most diversity, while NGS128, NGS130 and NGS131.H12 show the least diversity.

For each individual sample (data not shown), the four dog samples showed the lowest overall diversity across all four metrics. The samples with the most diversity tended to be a sheep and cow sample (NGS004 and NGS007) and a water sample (NGS126); however, there was some variability depending on the metric used, with three additional sheep and cow samples and two additional water samples sometimes in the top three samples. The different samples for each faecal source showed similar diversity.

3.4.2.3.2 β-diversity

Two-dimensional principal coordinates analysis (PCoA) plots were generated to determine the β-diversity between samples, using weighted and unweighted Unique Fraction metric (UniFrac) measures for discrete data analysis (Figure 3.7). The two water samples below the sampling depth threshold chosen are not depicted. Using a sampling depth threshold of 92 was evaluated to determine the effects of removing these two samples on clustering. The data points were found to be less tightly clustered overall, but still showed general source-level clustering (data not shown), particularly with the water samples.

UniFrac is based on the assumption that communities that differ more should require more unique evolution of the lineages they contain, through measurements of phylogenetic distance (Lozupone and Knight, 2005). By determining a UniFrac value for all pairs of multiple environments, a distance matrix can be produced, which can be used to cluster the environments using a hierarchical clustering algorithm, such as Unweighted Pair Group Method with Arithmetic mean (UPGMA), or to perform dimensionality reduction using PCoA (Liu *et al.*, 2007). Unweighted UniFrac is a qualitative measure, where duplicate sequences contribute no additional branch length, while weighted UniFrac is a quantitative measure as it detects changes in how many sequences from each lineage are present, as well as changes in which taxa are present (Lozupone and Knight, 2008). Both unweighted and weighted UniFrac measures provided similar plots, with samples from the same species generally clustering together. The unweighted data provided slightly tighter clustering for samples, however, accounted for only 14% of the variance, compared with 34% for the weighted analysis. The PC1 vs. PC2 plots also showed tighter clustering compared with the PC1 vs. PC3 plots, with the PC1 vs. PC3 plot for weighted analysis having the least clustering effects. In all cases, the water samples clustered together, with the sewage samples and the swan sample generally showing the most similarity.

3.4.2.3.3 Jackknifed support

A bootstrapped tree was generated for both the weighted and unweighted UniFrac data, which provides support for the PCoA sample clustering (Figure 3.8). While water samples NGS018-B4.18 and NGS128.H29 are not shown, due to having less OTUs than the sampling depth used, they were found to cluster with the rest of the water samples when a small sampling threshold was used (data not shown). NGS018.B4.18 was always found to cluster tightly with NGS018.H7.

Unweighted UniFrac provides greater support for the clustering effects, with all samples from each source clustering together. The least support is seen within the cow and sheep clusters, with bootstrap values for some of these nodes quite low, however, there is 100% support for all alpaca, cow and sheep samples to be clustered together. The water samples are also clustered together, although the sewage samples are included within this node. The water sample nodes are generally not as well supported as the source-specific nodes.

Weighted UniFrac does not show as much support for source-specific clustering, with the dog samples the most notably distinct, scattered throughout the tree. The ruminant species are still clustered together, but the majority of the other species are split up to some degree. The water samples are still clustered with the sewage samples, and show a slightly higher level of support.

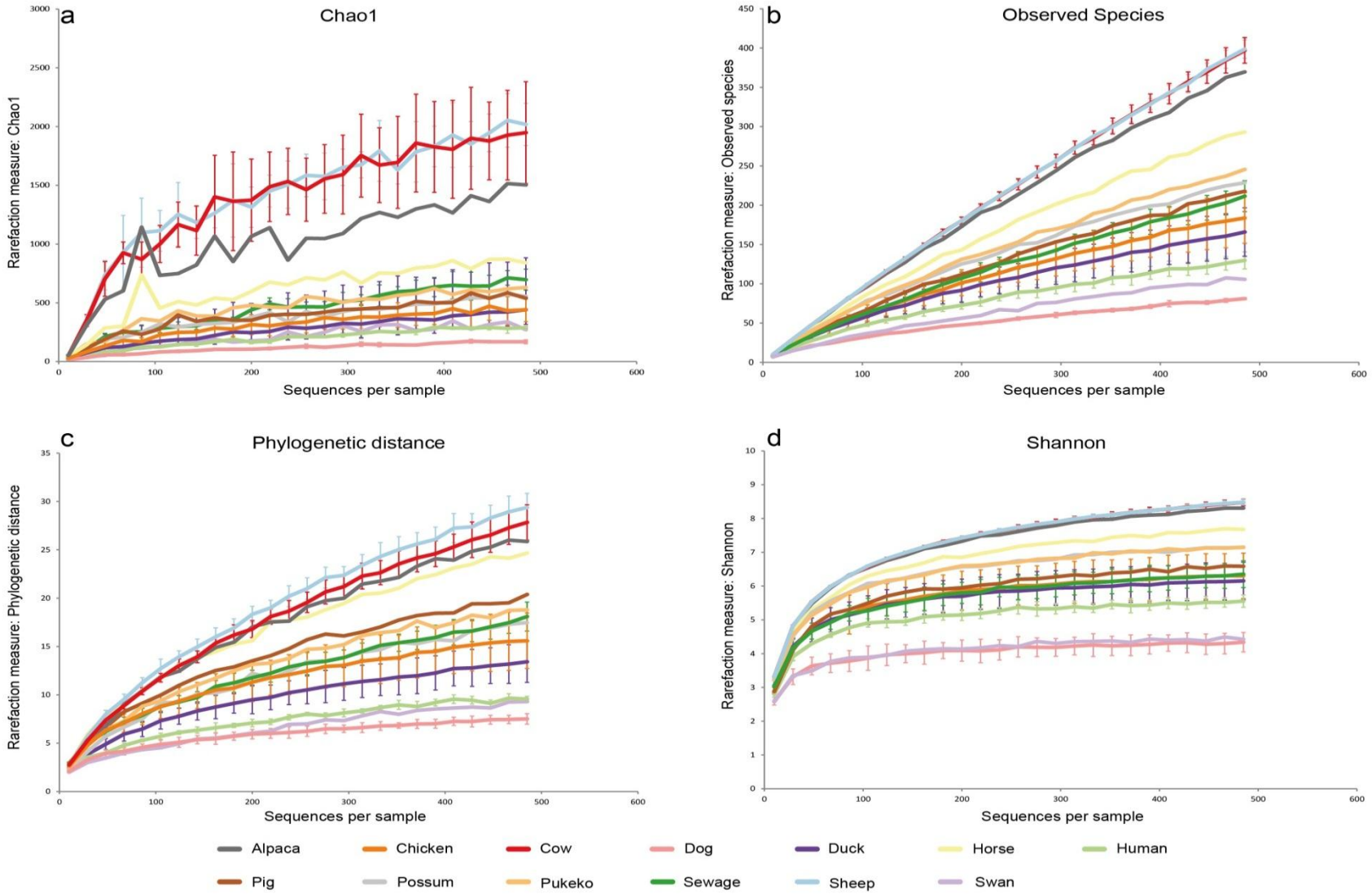


Figure 3.5: Rarefaction curves for faecal source library samples. Four alpha diversity metrics were used: Chao1 (a), Observed species (b), Phylogenetic Distance (c) and Shannon (d). For species where more than one sample was analysed, the average is plotted with error bars displaying the sample variation**.**

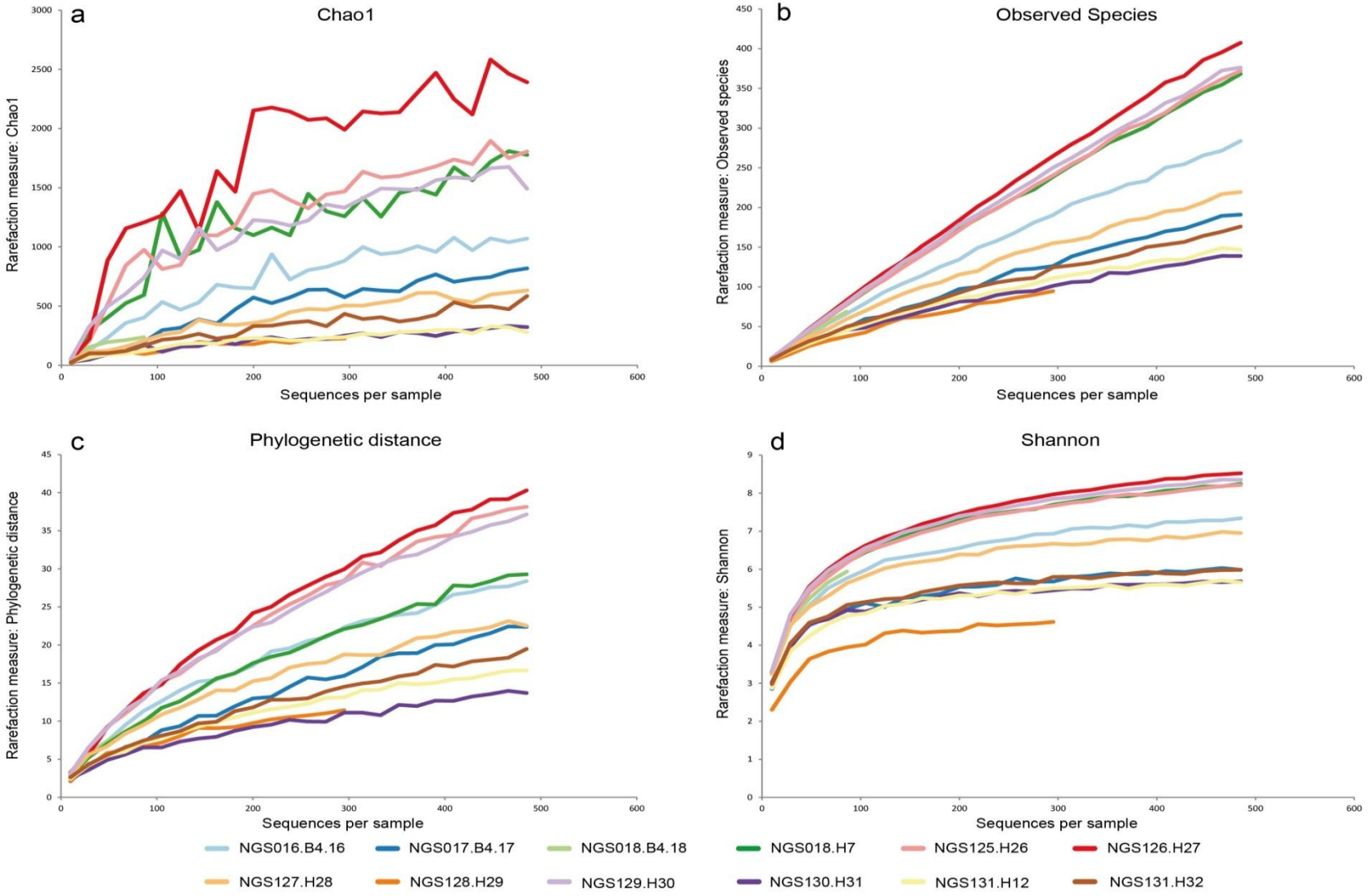


Figure 3.6: Rarefaction curves for water samples. Four alpha diversity metrics were used: Chao1 (a), Observed species (b), Phylogenetic Distance (c) and Shannon (d).

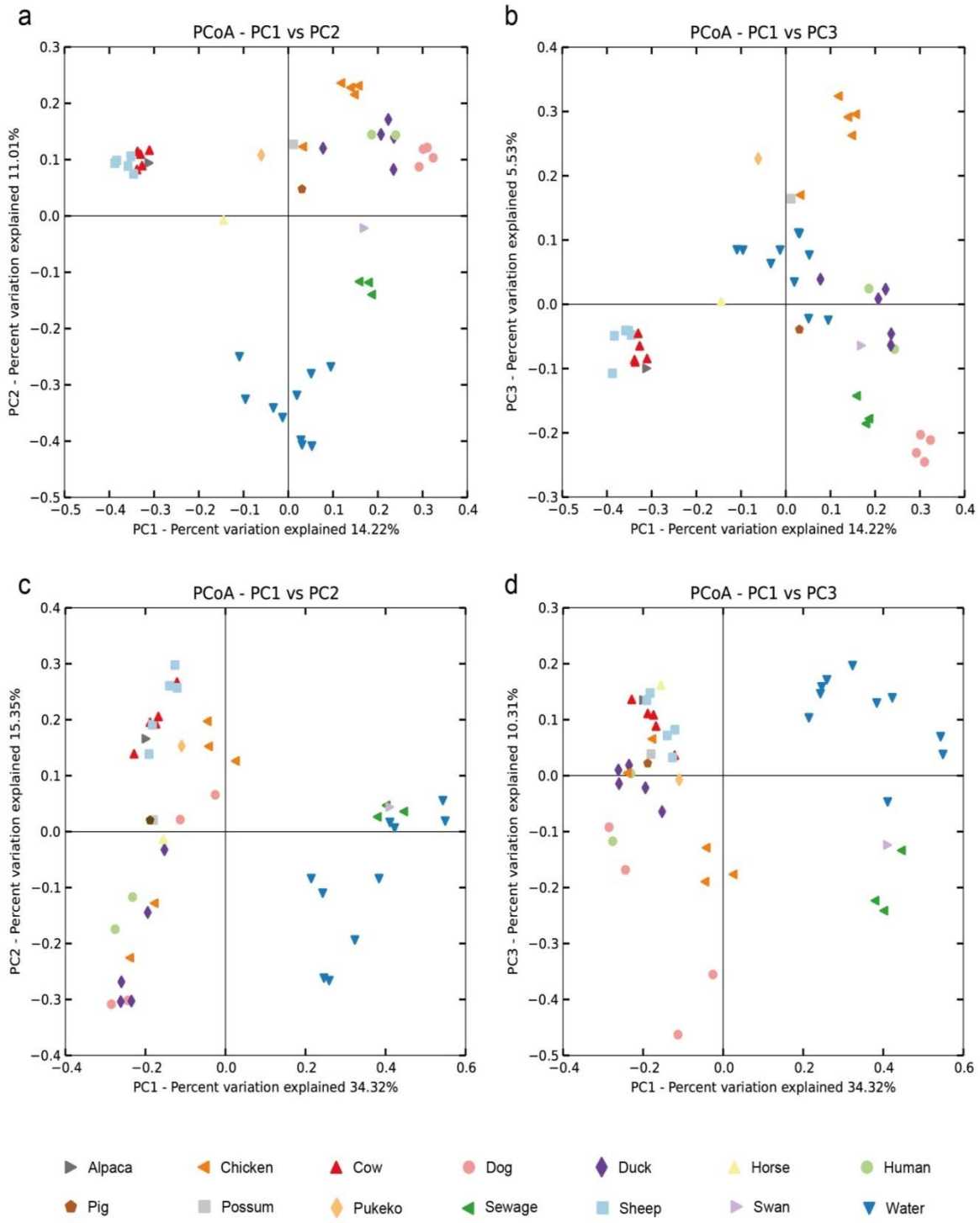


Figure 3.7: Two-dimensional PCoA UniFrac plots for total bacteria; unweighted PCoA for principal coordinates 1 vs. 2 (a) and 1 vs. 3 (b); weighted PCoA for principal coordinates 1 vs. 2 (c) and 1 vs. 3 (d).

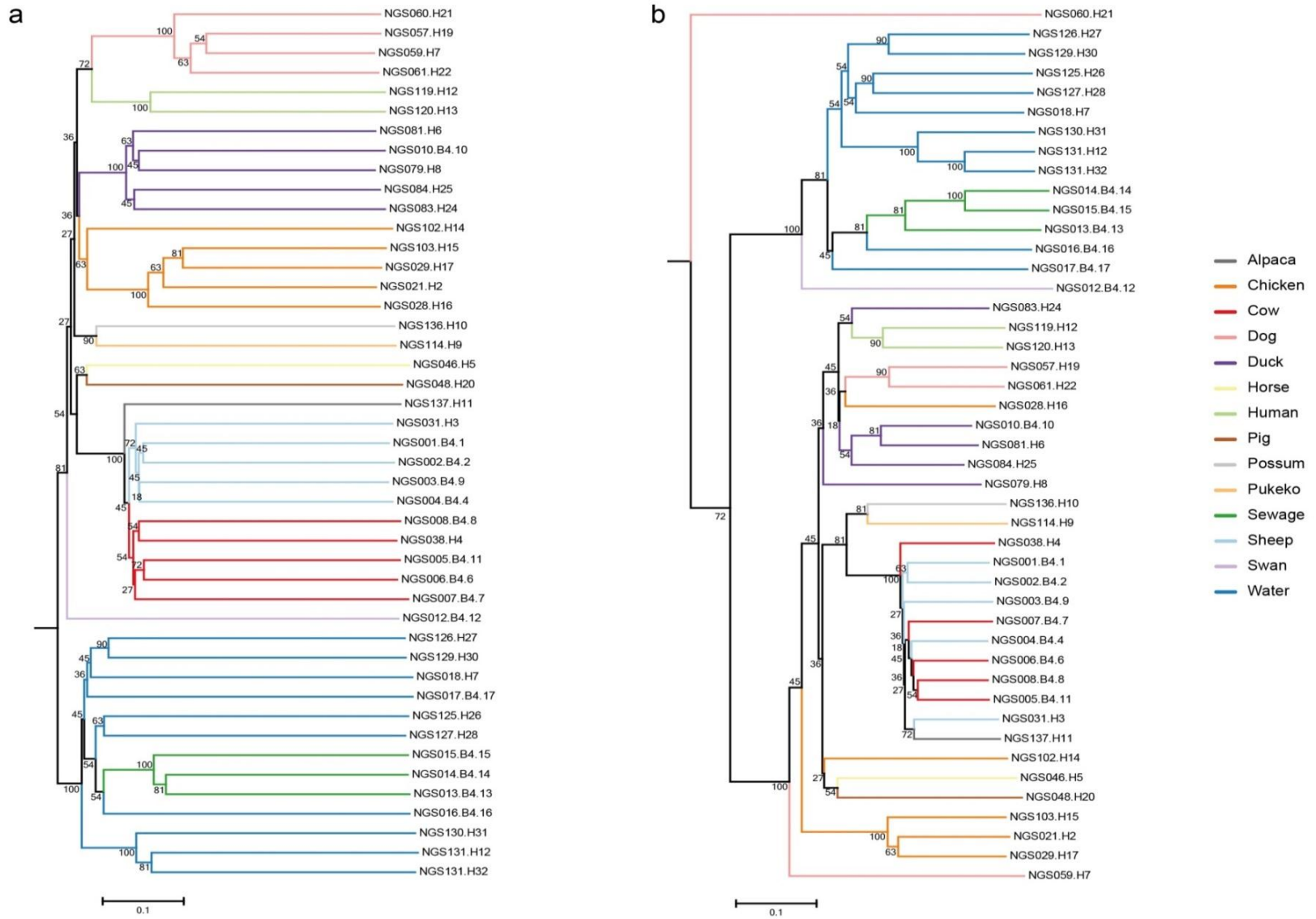


Figure 3.8: Jackknifed UPGMA bootstrapped trees for total bacteria. Unweighted UniFrac data clustering (a) and weighted UniFrac data clustering (b).

3.4.2.4 Faecal bacteria analysis

Bacteria phyla which are commonly found in environmental samples were filtered out of the data, to allow for community diversity analysis to be performed on faecal-associated bacteria only, including *Bacteriodetes*, *Fibrobacteres*, *Firmicutes*, *Fusobacteria*, and certain classes of *Proteobacteria* (*Alphaproteobacteria* and *Gammaproteobacteria*). The OTU data for this subset is provided in Table 3.6. The β-diversity PCoA and jackknifed diversity analyses were repeated with this subset of data (Figures 3.9 and 3.10). To ensure all water samples were included in the analyses, a sampling depth of 44 and 30 was used for PCoA and jackknifed β-diversity, respectively.

By analysing only the faecal bacteria, there were noticeable differences in the PCoA plots (Figure 3.9) compared with those for total bacteria (Figure 3.7). Less variation is explained for each of the principal coordinates, and the clustering of the different sources is not as pronounced, although samples from each source are still noticeably clustered together. There is still no suggestion of individual water samples clustering closely to their contamination source.

The unweighted bootstrapped tree (Figure 3.10a) is very similar to that for total bacteria (Figure 3.8a) with only one chicken sample (NGS102) showing a large change in clustering. However, the bootstrapped support for the majority of the nodes is not as strong compared to those for total bacteria. The weighted tree (Figure 3.10b) also shows a few differences from that generated for total bacteria (Figure 3.8b). The sewage samples only cluster with a single water sample, compared to being within the same clade for total bacteria. With the exception of NGS060, the dog samples all cluster together when only looking at faecal bacteria, compared to being spread out throughout the tree for total bacteria.

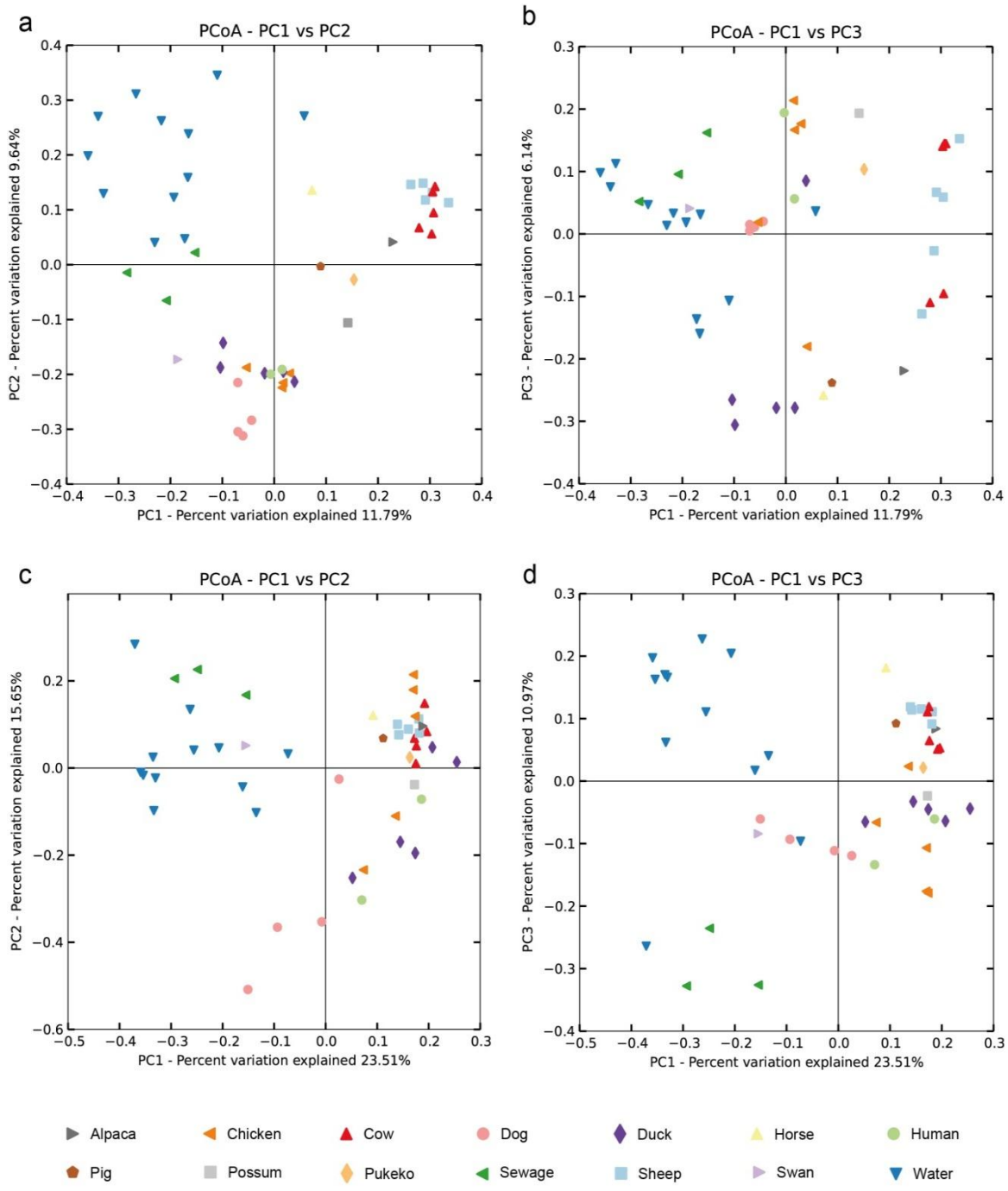


Figure 3.9: Two-dimensional PCoA UniFrac plots for faecal bacteria; unweighted PCoA for principal coordinates 1 vs. 2 (a) and 1 vs. 3 (b); weighted PCoA for principal coordinates 1 vs. 2 (c) and 1 vs. 3 (d).

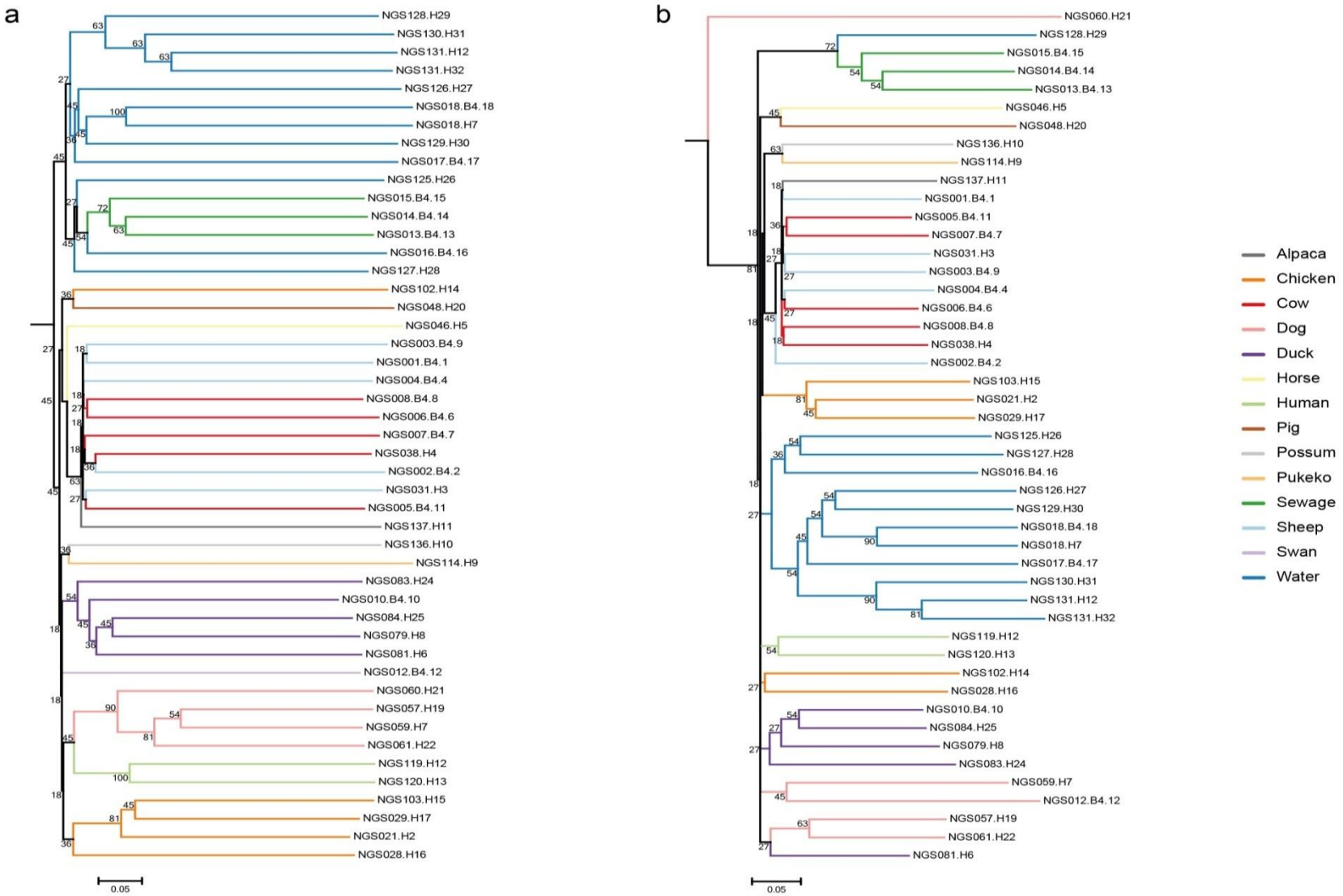


Figure 3.10: Jackknifed UPGMA bootstrapped trees for faecal bacteria. Unweighted UniFrac data clustering (a) and weighted UniFrac data clustering (b).

3.4.2.5 Sample diversity

The intra-species diversity can be examined closer by looking at the species that make up each sample sequenced from a single species. It has previously been found that identical samples that are processed in an identical manner do not always yield similar sequence data, with differences in bacterial community composition between individuals often quite large (Zhou *et al.*, 2011). There were six key species with multiple samples included in this study: chicken, cow, dog, duck, human and sheep. For each of these species, five composite samples were prepared, using DNA extracted from five individual faecal samples, with the exception of dog, for which only four composite samples were included. The five human samples contained two composite human faecal samples and three raw sewage samples, which are considered to represent a community population already, with no need for further pooling. Data from each individual sample from these species was analysed for the proportion of sequences of each genera identified in Table 3.7, to determine the variation between samples. Note that because these are composite samples, the comparisons are being made across different community subsets, and not individuals.

3.4.2.5.1 Bacteroides

Bacteroides were found to be present in all faecal libraries and in all individual samples, with relatively consistent percentages in chicken, cow, human, sewage and sheep samples. Of the four dog samples analysed, two were found to contain high percentages of *Bacteroides* (33 and 47%), while the other two samples contained only low percentages (4 and 7%). The five duck samples also varied, ranging from 12 – 37% (average 24%).

3.4.2.5.2 Prevotella

Prevotella was also found in almost all faecal libraries, with noteworthy average percentages in dog, duck, human and sewage libraries. Of these, *Prevotella* was only consistently found in the three sewage samples. The dog samples ranged from 0.5 – 77% (average 26%), duck samples ranged from 3 – 26% (average 14%) and only one human sample contained 15% (average 2%).

3.4.2.5.3 Ruminococcus and *5-7N15*

Ruminococcus was found in all faecal libraries, with noteworthy percentages in the ruminant species. The percentage values were relatively consistent across all samples

for each of the species it was found in, with much higher values in cow (17 – 28%) and sheep (16 – 22%) samples compared with chicken, duck and human samples (0.3 – 2%). *5-7N15* was also relatively consistent across the ruminant species, although present at a much lower level than *Ruminococcus*, ranging from 2 – 7%.

3.4.2.5.4 Sarcina, Megamonas and *J2-29*

Sarcina was identified as a potential dog specific marker, however, was only found to be greater than 1% in one dog sample (16%). *Megamonas* was also found to vary in percentages, from 3 – 16% (average 11%), but was found in all samples. *J2-29* ranged from 0.2 – 20%. Interestingly, the same dog sample contained the highest percentage of all three of these genera (NGS059).

3.4.2.5.5 Blautia, Roseburia and *Faecalibacterium*

Blautia, *Roseburia* and *Faecalibacterium* were identified as potential human markers, and *Roseburia* and *Faecalbacterium* were found to be reasonably consistent between the two human samples. *Blautia* was found to have only 1% in NGS119 and 7% in NGS120. All three of these genera were found in low levels in all the sewage samples.

3.4.2.5.6 Acidovorax, Zoogloea and *Arcobacter*

Acidovorax was only found in high numbers in the three sewage samples, ranging from 2 – 12% (average 7%). *Zoogloea* was also found in all three sewage samples, with two samples containing less than 1% and one containing 3%. *Arcobacter* provided the greatest percentage of all sewage samples, ranging from 15 – 42% (average 31%), which suggests that this could be a very promising sewage-associated marker.

3.5 Discussion

3.5.1 Taxonomy classifications

To define the taxonomy of 39,112 OTUs identified, a cut off at 97% sequence similarity was used, which is generally accepted as representing a species. This value is standard across many microbial studies, although is considered to be conservative by many (Mizrahi-Man *et al.*, 2013). 6% of the OTUs were classified as not bacterial, with a further 1% classified as unknown bacterial sequences.

3.5.1.1 Phyla taxonomic classification

There were 49 different phyla represented across the 13 source libraries and 12 water samples, with 27 of these considered to be candidate phyla divisions (McDonald *et al.*, 2012). The candidate divisions were predominantly found only in water samples, with AD3, OP11, SR1 and TM7 the only ones found within other sources. All phyla were represented within the collection of water samples, with faecal source libraries ranging from 5 phyla (human) to 23 phyla (sewage), with an average of 12.8 phyla per source. This range of diversity is similar to other faecal source studies using NGS techniques (Jeong *et al.*, 2011; Lee *et al.*, 2011; Unno *et al.*, 2010). Sewage is known to contain environmental bacteria as well as faecal-associated bacteria, which can account for the higher diversity seen.

Four phyla were found in all samples, *Bacteroidetes* (36% of all sequences), *Firmicutes* (32%), *Proteobacteria* (15%) and *Tenericutes* (2%), making up 85% of all sequences. Fusobacteria were only found in eight of the faecal source libraries and some of the water samples, making up a further 3% of the total sequences. The remainder 44 phyla accounted for only 5% of the total number of sequences.

In a similar study in South Korea, Jeong *et al.* (2011) found 19 different phyla represented across cow, human, pig and river water samples, with *Bacteroidetes* and *Firmicutes* accounting for 37-57% and 38-54% of bacterial sequences in the faecal samples, respectively. These values are slightly higher than those seen in this study; however, this may be due to differences in taxonomic classifications, as Jeong *et al.* used an 80% confidence threshold for the RDP Classifier. *Proteobacteria* were also found at similar levels in cow, human and pig faecal samples, as well as being the dominant phyla in water samples.

Direct comparison of the taxonomic phyla classifications for the different faecal sources (Figure 3.3) shows that most species have similar overall compositions. Chicken, sewage and swan sources are noticeably different for their higher *Proteobacteria* composition, while dog and swan samples stand out for having higher percentages of *Fusobacteria*. Sanapareddy *et al.* (2009) also found high levels of *Proteobacteria* in sewage, with the majority of sequences matching those previously found in water, soil and other wastewater studies, suggesting that most of the *Proteobacteria* found are likely to be environmental sources.

Comparison of the different water samples (Figure 3.4) also shows quite similar overall composition, with the majority of samples highly dominated by *Proteobacteria*. These results suggest that it is likely to be difficult to define potential faecal contamination from the phyla level alone, and more taxonomic information, such as genus or species level classification is needed.

3.5.1.2 Genus taxonomic classification

The genus level classifications were analysed to determine whether this taxonomic level provides enough diversity amongst sources to support faecal source tracking identification. 419 OTUs were classified to the genera level, of which 21 were identified as potential source markers, based on making up at least 5% of the total number of sequences in an individual source library (Table 3.7). The 5% threshold was chosen as an indication that the OTU is of the source origin, and not likely to be from the environment. It also allows for potential dilution in water samples, as lower percentages may become too low to be picked up from environmental water samples.

3.5.1.2.1 Bacteroidetes

Five *Bacteroidetes* genera were identified as potential source markers, including two from the *Bacteroidaceae* family, one from the *Porphyromonadaceae* family, one from the *Prevotellaceae* family and one from the *Flavobacteriaceae* family. Of these, the *Bacteroides* genus was the most represented, with eight faecal sources containing at least 5%. The *Bacteroides* genus is the most common *Bacteroidetes* target for MST markers, although is often clustered with *Prevotella* (Bernhard and Field, 2000a; Bernhard and Field, 2000b; Dick *et al.*, 2005b; Okabe *et al.*, 2007). However, while *Bacteroides* may make a strong marker for general faecal contamination, because so many hosts contain similar or identical sequences, this genus may not provide useful targets for determining source contribution (Dick *et al.*, 2005a).

The genus *Prevotella* was found in dog (25.7%), duck (14.3%) and pig (13.0%) samples, with only the pukeko sample not containing any sequences from this genus. *Prevotella spp.* have been used as a pig-specific target (Dick *et al.*, 2005a; Okabe *et al.*, 2007), however, the data here suggests that there is a high chance of cross-contamination of these markers with dog and duck samples.

Parabacteroides was found all but three faecal sources, with human samples the only source to contain more than 7%, suggesting this genus may be a potential human-specific MST target. Possum faecal samples contained the next highest overall percentage for this genus (4%). Possum faecal samples have previously been found to amplify with human-specific *Bacteroidales* PCR targets (Devane *et al.*, 2013). The *Bacteroidaceae* genera *5-7N15* cannot be associated to a single species, but appears higher in ruminant animals, so may be a potential ruminant marker.

3.5.1.2.2 Firmicutes

Nine *Firmicutes* genera were identified, representing six different families. Of these, seven genera were found at 5% or more in only one faecal source, and representing a range of faecal sources. Many of these are found in the gut of a range of animals, and are often associated with the breakdown of plant material. In a recent study of a range of faecal sources, Lee *et al.* (2011) found *Firmicutes* to be the primary phylum responsible for clustering, suggesting that bacteria within this phylum may provide useful MST targets.

Ruminococcus was found in five faecal sources, including alpaca, cow, possum, pukeko and sheep. *Ruminococcus spp*. are generally found in the gut of ruminants (Dowd *et al.*, 2008) but a number of species have also been found in human faecal samples (Wang *et al.*, 2004). The results from this study suggest that while this genera is found in a range of source species, it is present at much higher levels in ruminants, suggesting that it can be used as a ruminant marker, although may require additional studies to determine a suitable detection limit.

Three of the genera identified were found predominantly in human samples. *Faecalibacterium* contains a single species, *F. prausnitzii*, and Zheng *et al.* (2009) have recently developed a *Faecalibacterium* targeted assay which appears to be specific for human faecal and sewage samples. The study included samples from China, France and the USA, suggesting the marker may be a useful target that appears to be distributed globally.

Lactobacillus makes up over 20% of the sequences within pig samples, and has recently been used as a pig-specific target (Marti *et al.*, 2010). Unlike *Prevotella spp*. which are also used as pig-specific targets, there is no significant occurrence of this genus in the

other faecal sources studied, suggesting the *Lactobacillus* target may provide better source-specific results.

11% of the sequences from dog faecal samples were classified as *Megamonas*, which is commonly found in dog faeces (Beloshapka *et al.*, 2013). Jeong *et al.* (2011) found a small percentage of *Megamonas* in human samples, but did not include dog samples in the analysis. An additional *Clostridiales* genus, *Sarcina*, was also only found significantly in dog faecal samples, however, this genera has also been found to be abundant in cattle fed on corn (Durso *et al.* 2012).

3.5.1.2.3 Fusobacteria

Two key *Fusobacteria* genera were detected in the samples, with the swan sample dominated by *Cetobacterium*, and the dog samples by *J2-29*. *Cetobacterium spp*. have been found in human faeces (Finegold *et al.*, 2003), marine mammals (Foster *et al.*, 1995) and freshwater fish (Tsuchiya *et al.*, 2008). To date, there appears to be no studies undertaken on the microbial community of swans. Dog faecal samples have been shown to include a large number of *Fusobacteriaceae* genera (Beloshapka *et al.*, 2013; Suchodolski *et al.*, 2008).

3.5.1.2.4 Proteobacteria

Three *Betaproteobacteria* genera were detected, with *Sutterella* in chicken, and *Acidovorax* and *Zoogloea* in sewage samples. *Betaproteobacteria* are generally found throughout a range of environmental samples, so the use of these as potential MST markers must be critically examined. *Sutterella* has been found in human (Engberg *et al.*, 2000) and dog faeces (Greetham *et al.*, 2004), and *Acidovorax* is commonly found in activated sludge in wastewater treatment plants (Heylen *et al.*, 2008; Sanapareddy *et al.*, 2009). *Zooglea* was only found in the sewage samples, although only represented 1.4%. *Zooglea* bacteria are known to play an important part in sewage treatment (Rossello-Mora *et al.*, 1995), which may suggest this genera could be a marker for treated effluent if high enough levels can be detected.

3.5.1.2.5 Comparisons to water samples

Application of a 1% threshold to selected genera allowed six of the ten water samples to be classified as being predominantly contaminated by either human or ruminant sources (Table 3.9). The assignments of the potential contamination sources were compared to the outcome of commercial analysis performed by ESR using a range of MST

techniques, including faecal sterol analysis and PCR assays. All six water samples were assigned to the same contamination source as indicated by the analyses performed by ESR. As no duck-specific genera were identified, NGS125 could only be classified as having human contamination. Two of the three potential dog markers were present in NGS127, but at levels below 1%.

Only genera for ruminant or human sources were able to be used as source-specific markers in the water samples analysed. While some of the other identified genera could be applied to specific source contamination, including duck, dog, horse, pig and swan, they were not present in high enough amounts to be able to be used as a contamination source marker for any of the water samples.

3.5.1.2.6 Comparisons to other studies

Lee *et al.* (2011) recently conducted a NGS study on faecal samples, and suggested a number of potential genera which could be used as source-specific markers. Interestingly, none of their suggested markers were found in the samples in this study, including *Bifidobacterium* in humans, *Yania* in chickens and *Marinicola* in pig.

Jeong *et al.* (2011) also suggest a range of potential source-specific markers following NGS analysis on faecal and water samples. They suggest five different human-specific species from the *Bacteroidales* order, plus an additional five species from other phyla, including *Megamonas funiformis* and *Ruminococcus lactaris*, which are from genera identified as potential markers for dog and ruminant sources in this study. Jeong *et al.* also identify a *Lactobacillus* species as a potential pig-specific indicator.

Unno *et al.* (2010) used a density ratio of OTUs shared between a range of faecal and water samples to try to determine contamination sources. They determined that *F. prausnitzii* may be a strong human faecal source marker, which supports the findings of this study. Unno *et al.* suggest that due to the number of shared OTUs found across different sources, the use of a single genus for source identification may not be appropriate for MST.

All three of the above studies were conducted in South Korea, and the differences in potential source-specific genera compared to those identified in this study suggest there is evidence for geographical differences in the microbial communities within animal species across the globe.

3.5.2 α-diversity

3.5.2.1 Rarefaction curves

Divergence-based methods do not assume that all species are equally related to each other, and consider a community more diverse if the individuals are highly divergent from each other (Lozupone and Knight, 2008). Only one divergence-based method was included in this study (Phylogenetic distance). That this rarefaction plot is similar to the other three species-based indices (Figure 3.5) suggests that the OTUs being estimated are phylogenetically similar. The Shannon index provides the best-case rarefaction curve for all samples, with each curve beginning to plateau early in the sampling, while the Observed species index suggests that there is still a lot more diversity to be sampled in each of the species. This demonstrates the effect the choice of diversity index can have on the data. Previous studies have also shown how alpha diversity estimates can be affected by the methods used in preparing the samples, suggesting the importance of keeping methods as consistent as possible when comparing across multiple samples (Flores *et al.*, 2012).

Bird and human faecal samples have previously been found to show low species richness and diversity (Unno *et al.*, 2010), which matches results seen here, with chicken, duck, pukeko and swan all showing lower diversity compared with the other animals sampled. In addition, both the human faeces and human sewage show low diversity in comparison to other animal species, suggesting a lower variety of gut microbiota in humans.

The rarefaction curves generated for each of the water samples (Figure 3.6) suggest that there is bacterial diversity that has not been sampled. As many of these samples have quite low sequencing numbers, this strongly suggests that these sampling depths do not provide enough information on the diversity of the samples. Therefore, the sequences generated may not be fully representative of the actual diversity within these environments. The first step in overcoming this is to sequence the samples at a greater coverage, resulting in more sequences per sample. Inhibition due to substances co-extracted with the DNA may also be occurring during the PCR amplification steps, which may require further optimisation protocols to overcome, for example, through qPCR analysis of potential inhibitor solutions (Opel *et al.*, 2010).

3.5.2.2 Intra-source diversity of specific genera

Intra-sample diversity was also examined by comparing the different samples within each source for bacterial composition. Most of the genera identified in Table 3.7 were found in each sample for the relevant sources, although with varying quantities. The largest variation between samples was seen in the four dog samples. Five genera were identified that represented more than 5% of the total number of sequences. Of these, *Bacteroides* was the most consistent across the samples, ranging from 33 – 47%. *Prevotella* showed the largest variation between the four samples, with 0.5 – 77%. *Megamonas, Sarcina* and *J2-29* were all only found in high quantities in one dog, suggesting a large variation in different canine individuals. A recent study also found that dogs showed a much larger diversity between individuals, compared with cats (Handl *et al.*, 2011). The ruminant samples proved the most consistent, with both *Ruminococcus* and *5-7N15* found at consistent levels across all samples. These results suggest that different animal species may show large levels of variation between individuals for various bacteria species. Therefore, care must be taken when investigating potential source-specific sequences, as they may not be highly represented in every individual.

3.5.2.3 Water sample replicates

Two water samples were sequenced twice to determine the effect of sequencing on sample diversity. NGS018 was included in GS454-01 (NGS018.B4.18), and did not result in a large number of sequences, with only 92 OTUs identified. Consequently, this sample was also included in GS454-03 (NGS018.H7), which resulted in 9,556 OTUs. The two sequencing runs were performed by different providers (NZGL and Macrogen Inc.), using different barcode sequences and slightly different amplification protocols, due to different polymerase enzymes. All phyla classified for NGS018.B4.18 were found in NGS018.H7 in much greater numbers, plus an additional 15 phyla and 9 candidate divisions. NGS018.H7 was found to have a higher proportion of *Bacteroidetes* sequences than NGS018.B4.18, but lower *Proteobacteria* and Candidate divisions (Figure 3.4). This provides evidence for how sample preparation and sequencing procedure can affect the data obtained from NGS platforms. These two samples used the two different barcoded primer designed used for this study, and results may reflect differences in the amplification protocol used for the two sets of primers. However, NGS016 and NGS017 both have much higher sequencing read numbers,

which are comparable to read numbers obtained for the remainder of the water samples sequenced in GS454-03.

NGS131 was also sequenced twice, with each sample using the same amplification protocol and sequencing provider (Macrogen Inc., GS454-03). The data from these two samples was much more consistent, with the majority of phyla showing similar percentages between the two samples. NGS131.H12 contained three phyla/candidate divisions not represented in NGS131.H32, while NGS131.H32 contained only two additional phyla. However, these differences were negligible, representing only 10 individual sequences in NGS131.H12 and 2 in NGS131.H32. NGS131.H32 contained approximately 400 more sequences than NGS131.H12, and overall had a slightly lower percentage of *Bacteroidetes* and slightly more *Proteobacteria* compared with NGS131.H12 (Figure 3.4).

These results suggest that samples which are prepared and sequenced in the same manner, can easily be compared using these methods. However, caution must be taken when comparing samples which have different preparation protocols, as identified with the two NGS018 samples.

3.5.3 β-diversity

β-diversity provides a measure of the partitioning of biological diversity among different environments, defining the number of species shared between two environments (Lozupone *et al.*, 2007). As with α-diversity, a number of measures can be used to define β-diversity, which can be broadly divided into two categories, qualitative and quantitative measures. Qualitative measures use the presence/absence of data to compare community structure, while quantitative measures take the relative abundance of each organism into account. This study incorporated both a qualitative (unweighted UniFrac) and quantitative (weighted UniFrac) measure, as it has been shown that the two different measures can lead to dramatically different conclusions from the same data set (Lozupone *et al.*, 2007).

3.5.3.1 Principal Coordinate Analysis

β-diversity measures are often displayed using PCoA, which uses a distance matrix to plot the samples, based on the factors which describe as much of the variation as

possible. QIIME generates plots for the first three UniFrac principal coordinates (PC) which explain this variation, each resulting in slightly differing clustering effects. It has previously been demonstrated that both weighted and unweighted UniFrac measures have specific niches in the analysis of microbial communities, with different sample sets having clearer patterns of variation explained by one method or the other (Lozupone *et al.*, 2007). This suggests that using both types of measures can be critical to the understanding of the factors that underlie microbial diversity.

Microbes from a range of physical environments have been shown to cluster based on salinity, followed by similar type of environments, such as soils and sediments, and water samples (Lozupone and Knight, 2007). Ley *et al.* (2008a; 2008b) found that diet lead to the most pronounced clustering effect in a range of mammals and vertebrates, followed by hosts being from the same taxonomic order having similar bacterial communities. Differences between the vertebrate gut microbiota compared to those from free-living communities were also explored, looking at the distribution of different phyla, including *Bacteroidetes, Firmicutes* and *Proteobacteria*. Lozupone *et al.* (2007) analysed the effects of obesity and kinship on the microbial population of mouse gut microbiota, and found that unweighted UniFrac identified patterns of variation linking kinship, with mice clustering almost perfectly by mother, while weighted UniFrac gave no strong kinship associations, instead grouping most of the obese mice together in one cluster. Samples from different human body habitats have also been found to cluster closely based on body habitat, with high interpersonal variability within habitats, but low individual temporal variability (Costello *et al.*, 2009).

In this study, we have included plots for PC1 vs. PC2 and PC1 vs. PC3 for both unweighted and weighted UniFrac (Figure 3.7). Weighted and unweighted PCoA plots result in slightly different clustering of the data, but generally the same conclusions can be drawn: that microbial communities are dependent on host species. There are some differences between the two different PC plots, with PC1 vs. PC2 generally showing better clustering than PC1 vs. PC3. This is to be expected, as PC2 explains approximately 5% more of the variation found compared to PC3.

The three ruminant animals studied, alpaca, cow and sheep, all cluster tightly together for all analyses, suggesting that they have very similar bacterial communities, with lineages that share a common evolutionary history (Costello *et al.*, 2009). The dog and

chicken samples show the most diversity across the plots, suggesting a greater diversity between individuals within these species. All four dog samples contained the lowest amount of α-diversity, suggesting a lower number of different OTUs compared to the other faecal sources. However, the PCoA plots for weighted UniFrac suggest that the bacterial lineages between the four samples differ to a reasonable extent. This is likely to be explained by the large variations seen in the percentages of the main bacterial species found across the four dog samples.

3.5.3.2 Jackknifed support

Bacterial communities are usually too complex to sample completely, so jackknife replicates are often used to estimate the uncertainty in PCoA plots and hierarchical clustering of the communities. This technique regenerates the β-diversity using a subset of the sequences. 75% of the value used to generate α- and β- diversity is suggested, and therefore a value of 375 was used for this study. By default, QIIME generates 10 jackknife replicates of the available data, and uses UPGMA to generate bootstrapped trees which show how much support there is for the PCoA clustering. Nodes in the UPGMA cluster that are recovered in a large percentage of the jackknife replicates are considered robust to sampling effort (Lozupone *et al.*, 2007).

The weighted and unweighted UniFrac trees (Figure 3.8) generally support the clustering seen in the PCoA plots, with bootstrap values similar between the two methods. The unweighted tree (Figure 3.8b) shows better phylogenetic clustering of each species. In both methods, the water samples cluster with the human sewage, with NGS016 consistently related to the three sewage samples. This provides a strong indication that this water sample is contaminated with sewage, which is consistent with the analysis results found by ESR (Table 3.2). Beyond this, the water samples show no indication of sharing similar taxonomic composition to any of the faecal samples. This suggests that environmental phyla have a large influence on the water samples.

3.5.3.3 Faecal bacteria diversity

There are a number of phyla that are generally considered to be environmental, and are found extensively throughout a range of environments. As a way of removing the effects of these bacteria from the hierarchical clustering analysis, these phyla were removed from the analysis, resulting in a subset of sequences from phyla that are considered to be faecal bacteria. This resulted in PCoA plots (Figure 3.9) showing less

source-specific clustering, however, the water samples were still seen to mainly cluster by themselves. In general, sewage samples were in a similar area to the water samples, while the three ruminant sources were typically the furthest away. This suggests that while the removal of non-faecal bacteria had an impact on the clustering of samples, it did not remove all of the water-specific influence. An analysis using only *Bacteroidetes* bacteria was also tried (data not shown), but did not result in any major changes to the overall clustering of the samples.

3.5.4 Is amplicon sequencing quantitative?

The data presented here clearly shows that sequencing barcoded amplicon samples using 454 NGS methods does not always produce even depth across all samples. For each sequencing run, samples were diluted to the same concentration and pooled in equal amounts to produce an equimolar mixed sample. Theoretically, this should result in a similar number of sequences produced for each sample, which was clearly not the reality. It has been suggested that there will always be random processes involved during sequencing procedures, which will result in a Poisson distributed relative frequency of final products, however, a recent study looking at these effects determined the variation in coverage level between the amplicons is greater than would be expected due to random processes alone (Binladen *et al.*, 2007). Variability between and within sequencing centres, even when following the same procedures, has also been shown to occur (Schloss *et al.*, 2011).

3.5.4.1 Barcode bias

Bias towards certain barcode sequences during the amplification and sequencing procedures has been suggested as a potential issue, with studies finding certain sequences to be significantly overrepresented in final data. By pooling each PCR amplicon in an equimolar ratio, any PCR-associated bias should be removed; however, there is no direct ability to remove bias associated with the amplification steps involved in the 454 sequencing procedures. This could explain some of the differences within the data sets presented here. Based on total number of sequences for each sequencing run (section 3.4.1), the expected average number of reads per sample would be 10,600, 10,500 and 8,800 for GS454-01B, GS454-02 and GS454-03, respectively. Approximately 25% of each sequencing run was removed during the filtering steps,

resulting in approximate averages of 7,900 sequences per sample for GS454-01B and GS454-02, and 6,600 for GS454-03. Comparing these expected values to the combined sequences in table 3.6 provides an indication on which samples are overrepresented and which samples are underrepresented, with 6, 4 and 4 samples in the three NGS runs respectively, strongly overrepresented in the three sequencing runs, and 5, 6 and 3 strongly underrepresented. Interestingly, NGS018, which was very strongly underrepresented in the GS454-01B run, was overrepresented in the GS454-03 run.

There were also two barcode sequences which were used twice, once in GS454-02 and once in GS454-03. The H7 barcode (NGS018 and NGS079) was found to be overrepresented in both sequencing runs, while the H12 barcode (NGS119 and NGS131) was found to be underrepresented, suggesting that the barcode sequence used may influence the proportion of sequences produced.

3.5.4.2 Spiked samples

It has been suggested that the use of mock communities may assist with understanding the sequencing effects when undertaking sequencing, which is particularly important for amplicon sequencing, as sequence reads are not assembled into larger genomes. Therefore, any sequencing error may cause the sequence to be classified as a novel bacterium (Schloss *et al.*, 2011). Amend *et al*. (2010) looked at the relationship between read abundance and biological abundance by spiking a house dust sample with known quantities and identities of fungi. One order of magnitude difference was found in the read abundance among species that were known to be present in equal quantities. Schloss *et al*. (2011) created a mock microbial community, containing 21 strains of Bacteria and Archaea, representing a range of phyla, classes, orders, families and genera. The study identified the effects of artefacts generated by PCR and sequencing, and suggests that the inclusion of a mock community sample with each sequencing run would be of benefit for further analysis, allowing calculation of the rate of chimeraism, sequencing error rate and drift in representation of community structure (Schloss *et al.*, 2011). Zhou *et al*. (2011) studied spiked PCR amplicon samples and also concluded that amplicon sequencing is not quantitative, suggesting that caution is required when making quantitative inferences about β-diversity.

3.5.4.3 16S rRNA gene copy number

The number of genes coding for 16S rRNA in a bacterial genome can vary from one to 15 (Case *et al.*, 2007; Farrelly *et al.*, 1995), with a recent study finding 460 copies of the 16S rRNA gene from a sample set of 111 bacterial genomes (Case *et al.*, 2007). These intragenomic copies were also found to often differ in sequence, with 62% of bacteria with more than one 16S rRNA gene copy found to display some degree of heterogeneity. This variability of the 16S rRNA gene can have large consequences for microbial diversity analyses, which need to be taken into account if quantitative conclusions are to be drawn. The differing copies can lead to multiple taxonomic classifications for a single organism, resulting in inflated diversity measurements and skewed abundance estimates (Větrovský and Baldrian, 2013).

Different approaches to account for this variation have started to be included in 16S rRNA analyses. It has been shown that information on 16S rRNA copy numbers and genome sizes of genome-sequenced bacteria can be used as an estimate for the closest related taxon in an environmental sample, allowing estimates of relative abundance to be calculated for individual bacteria taxa (Větrovský and Baldrian, 2013). This can result in increases in the abundance estimates of some taxa and a decrease in others, depending on the number of 16S rRNA gene copies found previously in related genomes. A computational method has also been developed, which has been found to improve the ability to accurately measure diversity and abundance of communities (Kembel *et al.*, 2012).

3.6 Conclusions

This study uses next generation DNA sequencing as a tool to identify potential faecal source contamination in water samples. A barcoding strategy was used which allows for a large number of samples to be sequenced at once, making the most of the high-throughput abilities these sequencing platforms have to offer. A large range of faecal sources were analysed, resulting in sequence data from the 16S rRNA gene for a large range of bacteria found in each source. The taxonomy of each sequence was determined, allowing comparisons of the microbial community for each source to be made. This has resulted in a number of potential source-specific genera to be identified, which can potentially be included in the constantly growing toolkit for microbial source tracking. When these species were compared against ten water samples, six were able to

be linked to either ruminant or human sources of contamination, which was consistent with the results of previous MST methods. Increasing the depth of coverage for each water sample would allow for a greater confidence in assigning contamination sources based on the presence of source-specific markers.

Microbial community diversity was also analysed using a range of measurement methods. The α-diversity for each source suggested that the diversity was not completely sampled for any source, although the level of diversity differed quite dramatically between sources, with cow and sheep showing the largest amount of diversity, and dogs showing the least. The β-diversity suggested that bacterial communities are influenced most by their host, with all metrics used showing source-specific clustering. This was still apparent when only faecal-associated bacteria were analysed.

The water samples were also found to cluster together, regardless of the analysis method used. One sample was found to consistently cluster close to the three sewage samples, and this hierarchical clustering was supported by jackknifed bootstrap values, suggesting that NGS016 is contaminated by human sewage. This was also supported by previous MST methods, as well as the presence of identified source-specific genera, *Acidovorax* and *Arcobacter*.

PCoA methods proved useful for defining the diversity between different faecal sources, but were unable to provide conclusive evidence to suggest which contamination source was found in each water sample. Ideally, the water samples would have either clustered with the sewage and/or human faecal samples, or with the ruminant samples. PCoA analysis has previously been used successfully to identify salinity as the major environmental determinant in diverse physical environments (Lozupone and Knight, 2007). Human microbiota has also been shown to vary systematically across body habitats and time (Costello *et al.*, 2009). However, the data described in this study suggests that clustering of bacterial communities for faecal samples may not provide a suitable analysis tool for identifying faecal contamination in water samples. The use of source-specific bacterial targets amplified by PCR continues to be a promising direction for microbial source tracking. Next generation sequencing data is providing a rich source of novel bacterial sequences with the potential to identify source-specific MST markers.

Chapter Four
Interrogation of next generation sequencing data using published PCR assays

4.1 Abstract

The ability to identify sources of faecal pollution in waterways is important for minimising human health risks and reducing the impact faecal contamination has on the environment. *Bacteroidales* species are promising markers for microbial source tracking, with many published PCR assays targeting a range of source-specific species. Water bodies may be contaminated by more than one faecal source, which requires multiple PCR assays to determine all contributing sources. In this study, sequences generated by next generation sequencing techniques were analysed for a range of published PCR assay markers, using a computational nucleotide sequence motif-based search method. All samples contained the *Bacteroidales* AllBac 296F primer sequences which potentially could indicate general faecal pollution. Twenty primer sequences targeting *Bacteroidales* associated with ruminants, humans, dogs and pigs were screened against faecal libraries from 13 different source libraries. The number of sequences found was expressed as a percentage of the number of AllBac 296F marker sequences found. Four ruminant, six human, one dog and three pig markers were found to be source-specific when a 2% threshold was applied, with the exception of some cross-reactivity with chicken and possum sequences. This method was validated against ten water samples from various sources in New Zealand, with contamination sources of six samples accurately identified, based on previous analysis results. The remaining four samples were unable to be confidently assessed due to limitations in the number of sequences. These results suggest that next generation sequencing data can be used to rapidly determine the faecal contamination of water sources, based on currently published PCR assays.

4.2 Introduction

The water quality of many waterways and coastal waters is deteriorating due to faecal contamination from human and animal sources, which can introduce a wide range of bacterial, viral and protozoan pathogens, impacting human health (Shuval, 2003). The presence of faecal contamination has traditionally been monitored through the use of indicator bacteria, such as *E. coli*, *enterococci* and culturable coliforms (Tallon *et al.*, 2005). However, these organisms are found in a wide range of warm- and cold-blooded animals, which prevents the identification of the actual source of contamination (Field and Samadpour, 2007). Microbial source tracking (MST) is an increasingly used approach which aims to determine host-specific contributions of faecal contamination, which can aid in appropriate corrective measures being identified to eliminate the pollution and minimise the impact on human disease.

There have been numerous MST methods developed with the intention of being able to discriminate between different faecal sources. Initial studies focused on library-dependent methods, such as pulse-field gel electrophoresis (Casarez *et al.*, 2007a) and antibiotic resistance analysis (Booth *et al.*, 2003), which match genetic or phenotypic patterns of known faecal indicator bacteria isolates to those from environmental sources. Library-independent methods utilising genetic markers associated with particular animal faeces have largely replaced the library-dependent methods, as they do not require a large isolate library, and generally use much faster and cheaper methodologies. These methods generally utilise polymerase chain reaction (PCR) to amplify targeted source-specific markers to determine the presence of various faecal contamination sources. Other approaches have also been studied, including human-specific viruses (McQuaig and Noble, 2011), chemical (Hagedorn and Weisberg, 2009), community-based (Cao *et al.*, 2011) and metagenomic methods (Unno *et al.*, 2010).

The 16S rRNA gene is the most commonly used species proxy for MST genetic markers (Cardenas and Tiedje, 2008), as it is composed of a number of conserved regions interspaced with nine hypervariable regions (V1-V9) (Chakravorty *et al.*, 2007). The conserved regions allow universal primers to be designed which can target all bacterial species, while the hypervariable regions allow for taxonomic identification across all phyla. PCR primers designed for MST have targeted a range of these regions, with no region receiving universal acceptance. However, the V1-V3 region has been

suggested to provide a higher degree of classification accuracy (Kim *et al.*, 2011; Wang *et al.*, 2007) and lower levels of bias towards specific taxonomic groups (Vilo and Dong, 2012).

It has been established that certain *Bacteroides spp*. are frequently associated with human faeces but not with animal species (Kreader, 1995; Stoeckel and Harwood, 2007), and many of the host-specific MST assays use primers targeting this genera or higher levels of its taxonomy (Bernhard and Field, 2000a; Savichtcheva *et al.*, 2007). A comparison study by Griffith *et al.* (2003) showed *Bacteroidales* to be the most accurate MST method for discriminating between human and non-human impacts using tests of mixed faecal sources in aqueous samples. There have been a number of general assays developed, which target all *Bacteroidales* (Kildare *et al.*, 2007; Layton *et al.*, 2006; Okabe *et al.*, 2007; Siefring *et al.*, 2008). There is also a panel of host-associated assays, including human, ruminant, dog, pig, horse, elk and Canada geese (Bernhard and Field, 2000a; Bernhard and Field, 2000b; Dick *et al.*, 2005a; Dick *et al.*, 2005b; Fremaux *et al.*, 2010; Kildare *et al.*, 2007; Layton *et al.*, 2006; Mieszkin *et al.*, 2009; Mieszkin *et al.*, 2010; Okabe *et al.*, 2007; Reischer *et al.*, 2007; Reischer *et al.*, 2006; Stricker *et al.*, 2008). However, many of these have only been evaluated on a small number of candidate sources, limiting the ability to assess cross-reactivity.

Faecal contamination often occurs through mixed sources, requiring multiple assays to be used in order to determine all faecal sources. Concentration of various DNA marker sequences within a specific host have also been found to be variable, suggesting the need to use multiple marker assays to reliably detect inputs from certain individuals (Kildare *et al.*, 2007). This is often expensive and time-consuming, particularly if there are multiple samples to be tested. Therefore, most laboratories only screen for a limited number of key sources, which may result in contamination sources being missed. Since no PCR assay is completely selective as an indicator towards a single source, there is also the possibility of contamination being assigned to the wrong source.

The recent advances in next generation sequencing (NGS) technologies have resulted in the ability to rapidly sequence large amounts of DNA from multiple samples simultaneously, at increasingly reasonable cost. These techniques have typically been applied to metagenomic studies, where entire microbial communities are analysed from an environmental sample. Analysis programmes such as QIIME (Caporaso *et al.*, 2010)

and Mothur (Schloss *et al.*, 2009) look at each sequence individually, assigning taxonomic classifications and providing information on microbial community diversity. While useful, these analyses are often computationally challenging, time consuming and can result in a large volume of information which is not always needed.

Water quality guidelines around the world are based on the correlation of indicator bacteria such as *E. coli* and *enterococci* with health affects as determined through a range of epidemiological studies (Cabelli, 1983; Cheung *et al.*, 1990; McBride *et al.*, 1998). With the advent of PCR based assays, new epidemiological studies have been recently undertaken in the USA to determine the relationship between PCR-based detection and health outcomes. Repeating these types of studies to accommodate NGS approaches described in the previous chapter, will be expensive, and will take a number of years to complete. In the meantime strategies based on published PCR assay sequences may be somewhat comparable.

We propose a simple computational nucleotide sequence motif search-based method that allows for rapid determination of multiple contamination sources from environmental water samples, utilising the large sequence datasets generated by NGS. Targeted PCR amplicons are generated from each sample using universal bacterial primers, and sequenced using NGS platforms, such as the Roche 454 GS FLX. Sequences are then screened against a pool of *Bacteroidales* host-specific primer and probe sequences (motifs), using the Geneious bioinformatics software (BioMatters, New Zealand). The inclusion of a general *Bacteroidales* assay allows a probability ratio to be determined for each host-specific motif, removing the impact non-*Bacteroidales* sequences may have on the overall number of sequences.

In order to demonstrate the potential of this approach, we have analysed a range of faecal sources, including alpaca, chicken, cow, dog, duck, horse, human, pig, possum, pukeko, sewage, sheep and swan, as well as ten contaminated water samples from various sources in New Zealand. Published assays screened include a range of *Bacteroidales* targeted assays, including the general AllBac assay (Layton *et al.*, 2006), three ruminant-specific markers (Bernhard and Field, 2000b; Kildare *et al.*, 2007; Reischer *et al.*, 2006), three human-specific markers (Bernhard and Field, 2000b; Kildare *et al.*, 2007; Reischer *et al.*, 2007), two dog-specific markers (Dick *et al.*, 2005b; Kildare *et al.*, 2007) and one pig-specific marker (Mieszkin *et al.*, 2009).

4.3 Materials and methods

DNA sequences from the 35 faecal samples and 10 water samples prepared in the previous chapters (Tables 3.1 and 3.2) were analysed using Geneious (Biomatters, New Zealand). The water samples had previously been analysed by a range of MST techniques (faecal sterol analysis, PCR assays), with probable contamination sources assigned to each.

Raw NGS read files were imported into Geneious, and poor quality regions were removed if not previously done so by the sequencing provider. The barcode sequences used were imported into Geneious, and the DNA sequences sorted based on these barcodes. Samples which did not perfectly match a barcode sequence were discarded, and the barcode sequences were removed from each DNA sequence read, to ensure no interference for subsequent analyses. PCR primer sequences were left incorporated in each DNA sequence, as these may be shared by other primer motifs being screened.

Faecal source libraries were created by combining all sequences from each source together, producing 13 source libraries. A range of published MST 16S rRNA source-specific assays, containing at least one primer which targets within the V1-V3 region, were selected for analysis (Table 4.1). Primer and probe sequences were imputed into Geneious as oligonucleotide sequences (motifs), and screened against each faecal source library and water sample. Where reverse primers were beyond the Univ529R primer target, they were not included in the analysis. Motif sequences were screened against the library DNA sequences with no mismatches allowed.

The number of sequences with a match to each AllBac motif sequence was divided by the total number of sequences in each sample to determine the percentage of *Bacteroidales* sequences. The number of sequences for each *Bacteroidales*-targeting motif was divided by the AllBac 296F motif matches.

DNA sequences from the ten water samples were screened and analysed in the same manner.

Table 4.1: *Bacteroidales* 16S rRNA assays targeted in this study.

Assay name	Target	Forward sequence	Reverse sequence	Probe sequence	Reference
AllBac	Total *Bacteroidales*	AllBac 269F GAGAGGAAGGTCCCCCAC	AllBac 412R CGCTACTTGGCTGGTTCAG	AllBac 375R CCATTGACCAATATTCCTCACTGCTGCCT	Layton *et al.* (2006)
BacCow	Ruminant-specific *Bacteroidales*	CF128F CCAACYTTCCCGWTACTC	BacCow 305R GGACCGTGTCTCAGTTCCAGTG	BacCow 257p TAGGGGTTCTGAGAGGAAGGTCCCCC	Kildare *et al.* (2007)
BacR	Ruminant-specific *Bacteroidetes*	BacR F GCGTATCCAACCTTCCCG	BacR R CATCCCCATCCGTTACCG	BacR P CTTCCGAAAGGGAGATT	Reischer *et al.* (2006)
CF193	Ruminant-specific *Bacteroides*	CF193F TATGAAAGCTCCGGCC	Bac708R (not analysed) CAATCGGAGTTCTTCGTG		Bernhard and Field (2000b)
BacH	Human-specific *Bacteroidetes*	BacH F CTTGGCCAGCCTTCTGAAAG	BacH R CCCCATCGTCTACCGAAAATAC	BacH P-pC TCATGATCCCATCCTG BacH P-oT TCATGATGCCATCTTG	Reischer *et al.* (2007)
BacHum	Human-specific *Bacteroidales*	BacHum 160F TGAGTTCACATGTCCGCATGA	BacHum 241R CGTTACCCCGCCTACTATCTAATG	BacHum P TCCGGTAGACGATGGGGATGCGTT	Kildare *et al.* (2007)
HF183	Human-specific *Bacteroides*	HF183F ATCATGAGTTCACATGTCCG	Bac708R (not analysed) CAATCGGAGTTCTTCGTG		Bernhard and Field (2000b)
BacCan	Dog-specific *Bacteroidales*	BacCan 545F1 GGAGCGCAGACGGGTTTT	BacUni 690R1(not analysed) CAATCGGAGTTCTTCGTGATATCTA	BacUni 656p (not analysed) TGGTGTAGCGGTGAAA	Kildare *et al.* (2007)
DogBac	Dog-specific *Bacteroidales*	DF475F CGCTTGTATGTACCGGTACG	Bac708R (not analysed) CAATCGGAGTTCTTCGTG		Dick *et al.* (2005b)
Pig2Bac	Pig-specific *Bacteroidales*	Pig2Bac 41F GCATGAATTTAGCTTGCTAAATTTGAT	Pig2Bac 163Rm ACCTCATACGGTATTAATCCGC	Pig2Bac 113P TCCACGGGATAGCC	Mieszkin *et al.* (2009)

4.4 Results

4.4.1 Faecal library validation

The faecal libraries for each source were screened for the selected motif sequences, with no mismatches allowed in the primer binding site (Table 4.2).

4.4.1.1 General markers

The three AllBac general *Bacteroidales* sequence motifs were found in all faecal libraries, with the percentage of the total number of sequences ranging from 8 – 46%. Variation in the percentage was seen between the two primers and probe, with no one assay sequence found at a consistent percentage across all libraries. The AllBac 296F motif was selected for determining the percentage of each source-specific marker, based on the total number of *Bacteroidales* sequences identified by the AllBac 296F motif.

4.4.1.2 Ruminant markers

Of the seven ruminant-specific motif sequences tested, one did not find any matches in any of the faecal libraries (CF193F). The reverse and probe motifs for BacCow were found in high levels for all libraries. CF128F was found in all three ruminant libraries, alpaca, cow and sheep. Low levels were also found in chicken, duck, human, sewage and swan faecal libraries. The three BacR motifs were also found in the three ruminant sources, with the forward motif also found in the pig library, and the probe motif in the chicken library. Other low level library matches for these three motifs include dog, horse, human, sewage, duck and swan.

4.4.1.3 Human markers

The eight different human-specific nucleotide sequences motifs tested all showed some level of source specificity, with the exceptions of BacH R, which did not yield any matches within the faecal libraries, and BacHum R, which was found in all faecal sources except horse and pig. Human faecal samples were positive for seven human-specific motifs, while sewage was positive for four (BacH P-oT, BacHum F, BacHum R and HF183). Sewage was also found at low levels for two of the BacH motif sequences (BacH F and BacH P-pC). Cross-reactivity was evident with possum faecal samples for six motifs, but was not detected with BacH P-oT. Chicken also showed some level of cross-reactivity, with positive results for three motif sequences. BacH F, BacH P-pC,

BacHum H and HF183 were all only present in human, sewage and possum libraries. BacH P-oT and BacHum P were found at low levels in cow, sheep and dog faecal libraries. The BacH F, BacH P-pC, BacHum F and HF183 motif sequences show specificity to human and sewage faecal sources, but also show cross-reactivity to possum faecal sources. Results suggest that the BacHum P and BacH P-oT motif sequences are useful as human markers, but show a wider range of low level cross-reactivity. A multi-assay approach appears necessary to allow source contamination to be confirmed as most likely being from human origin.

4.4.1.4 Dog markers

The dog-specific marker BacCan 454F was not found in any of the faecal libraries, while the DogBac DF475F matched 33% of the *Bacteroidales* sequences in the dog samples. One sequence match was found in each of the cow, duck and sewage libraries. The high specificity to dog faeces suggests a low detection threshold should be suitable for detecting dog-specific contamination with this marker.

4.4.1.5 Pig markers

All three Pig2Bac markers were only found in the pig faecal sample, but only represented 4-5% of the total *Bacteroidales* sequences in the library. This suggests that these markers are highly specific to pig faeces, but may not be present in high numbers.

4.4.2 Determining a specificity threshold

In order to accurately apply this method to environmental water samples, a threshold for accepting a motif match as a potential contamination source needs to be set. Three motif sequences (CF193, BacH R and BacCan 454F) did not find any sequence match across all source libraries, so have been removed from further analysis. Another three motif sequences could not differentiate between different sources, so have been assigned as a non-specific source marker (BacCow 305R, BacCow 257P and BacHum R). The remainder of the motifs show a much higher level of specificity, suggesting that a threshold of 5% of the number of *Bacteroidales* sequences may provide a good starting point for assessing contamination sources. Table 4.3 summarises the specificity for each source-specific motif sequence found to match sequences within the faecal libraries, based on a requirement of 5%. There is always potential for errors to be made within

sequencing platforms, therefore, very low numbers of sequences are less likely to be indications of contamination. Setting a minimum detection threshold of 2% is likely to remove possible sequencing artefacts, while still allowing low level specificity to be noted.

4.4.3 Water sample validation

Ten faecal contaminated water samples were screened for the primers which were found in at least one faecal library source (Table 4.4). Between 339 and 11,113 sequences were obtained from each sample. NGS130 matched no source-specific motif sequences, and NGS128 only had a single sequence match for two markers, suggesting that samples with low sequence numbers are less likely to result in motif matches.

4.4.3.1 Non-specific markers

The six motifs which had low source specificity in the faecal libraries were found in almost all the water samples. The percentage of *Bacteroidales* sequences ranged from less than 1%, up to 45% of the total number of sequences.

4.4.3.2 Ruminant markers

CF128F was found in four of the five samples previously identified as containing ruminant contamination, with the fifth sample not matching any markers. BacR R and BacR P were also found in all four samples, with BacR P also matching sequencing in four of the human contaminated samples and BacR R found in one sample.

4.4.3.3 Human markers

BacHum P and BacH P-oT were both found in all five human contaminated water samples, with low levels also in two of the ruminant contaminated samples. BacH F, BacH P-pC, BacHum F and HF183 were all found in three of the five samples and none of the ruminant contaminated samples. NGS131 only matched very low numbers of sequences, probably due to a very low *Bacteroidales* sequence percentage overall.

4.4.3.4 Dog and pig markers

Only one sample (NGS127) contained the dog marker DF475F, with 3% of the *Bacteroidales* sequences matching the assay sequence. One sample (NGS018) contained two sequence matches for each of the Pig2Bac assay sequences, representing

less than 1% of the total *Bacteroidales* sequences, as all three markers were found in the same two sequences.

4.4.3.5 Assigning contamination sources

4.4.3.5.1 Removing poor quality samples

The total number of sequences available for analysis has some impact on the proportions assigned to each marker, limiting the confidence we can have in assigning a contamination source. Therefore, a minimum number of sequences for a sample should be determined. NGS128 has the smallest number of sequences (339), with a very low percentage of *Bacteroidales* sequences, as determined by AllBac 296F. Only two motifs have a single sequence match, suggesting we cannot assign this contamination with a high level of confidence. NGS016 and NGS017 also have a small number of sequences, with NGS016 having a higher percentage of *Bacteroidales*. This suggests that a minimum percentage of *Bacteroidales* sequences may also be beneficial. A minimum *Bacteroidales* threshold can then be applied to samples such as NGS130 and NGS131, which have higher numbers of overall sequences, but very little *Bacteroidales* sequences. This low proportion results in biases for calculations of each source-specific motif sequence. We propose using a low level sequence filter of 1,000 sequences, with a 2% *Bacteroidales* motif sequence threshold, using the AllBac 296F motif. Applying these thresholds to the water samples analysed here results in the removal of four samples, NGS017, NGS128, NGS130 and NGS131.

4.4.3.5.2 Assessing source-specific motifs

The same 2% motif sequence threshold suggested for *Bacteroidales* motifs can be applied to all the source-specific motifs. This allows potential contamination sources to be assigned to all six of the remaining water samples (Table 4.5), which match MST results previously obtained for these samples. Only one sample contains a positive result for the dog-specific motif sequence (NGS127), which correlates with the previous MST data. Two sequences which match the pig-specific motif sequences were found in one sample (NGS018). Although this equates to only 1% of the *Bacteroidales* sequences, as these motifs were only found in pig faecal samples, it may suggest a very low level of pig contamination in this water sample.

Table 4.2: Sequence numbers and percentages for *Bacteroidales*-specific motifs screened against faecal libraries. Percentages for AllBac motifs are of total sequences, with percentages of source-specific motifs of total *Bacteroidales* sequences, determined by AllBac 296F motif sequences.

	Ruminant animals				Non-ruminant animals			Human			Birds		
	Alpaca	**Cow**	**Sheep**	**Dog**	**Horse**	**Pig**	**Possum**	**Human**	**Sewage**	**Chicken**	**Duck**	**Pukeko**	**Swan**
Total sequences	17355	35474	70409	44964	6068	9523	7268	11659	57335	44450	53064	8907	11777
AllBac 296F	3824 (22%)	6743 (19%)	11813 (17%)	12628 (28%)	1243 (20%)	2426 (25%)	1418 (20%)	2220 (19%)	5091 (9%)	3550 (8%)	19900 (38%)	1381 (21%)	1241 (11%)
AllBac 412R	2138 (12%)	4247 (12%)	7788 (11%)	19385 (43%)	507 (8%)	1799 (31%)	865 (12%)	3877 (33%)	5086 (9%)	7679 (17%)	24396 (46%)	647 (8%)	1742 (15%)
AllBac 375P	4388 (25%)	10242 (29%)	17587 (25%)	12628 (28%)	2037 (34%)	2997 (31%)	2422 (33%)	4179 (36%)	9168 (16%)	6516 (15%)	23617 (45%)	1275 (14%)	1717 (15%)
BacCow CF128F	950 (25%)	1519 (23%)	2182 (18%)	-	-	-	-	5 (0%)	12 (0%)	101 (3%)	91 (0%)	-	2 (0%)
BacCow 305R	3238 (85%)	6156 (91%)	11447 (97%)	110 (1%)	446 (36%)	390 (16%)	44 (3%)	247 (11%)	19246 (378%)	3198 (90%)	2990 (15%)	753 (41%)	1072 (86%)
BacCow 257P	3749 (98%)	7728 (115%)	13435 (114%)	12881 (102%)	1204 (97%)	2383 (98%)	1411 (100%)	2248 (101%)	6123 (120%)	3582 (101%)	22028 (111%)	1818 (99%)	1707 (138%)
BacR F	1003 (26%)	1371 (20%)	2175 (18%)	1 (0%)	1 (0%)	212 (9%)	-	-	7 (0%)	2 (0%)	64 (0%)	-	3 (0%)
BacR R	912 (24%)	884 (13%)	906 (8%)	-	1 (0%)	-	-	5 (0%)	15 (0%)	100 (3%)	568 (3%)	-	-
BacR P	756 (20%)	1739 (26%)	2344 (20%)	278 (2%)	34 (3%)	34 (1%)	-	5 (0%)	81 (2%)	276 (8%)	79 (0%)	-	3 (0%)
CF193	-	-	-	-	-	-	-	-	-	-	-	-	-
BacH F	-	-	-	-	-	-	245 (17%)	1615 (73%)	219 (4%)	-	-	-	-
BacH R	-	-	-	-	-	-	-	-	-	-	-	-	-
BacH P-pC	-	-	-	-	-	-	250 (18%)	1651 (74%)	146 (3%)	-	-	-	-
BacH P-oT	-	2 (0%)	10 (0%)	36 (0%)	-	-	-	272 (12%)	992 (19%)	2458 (70%)	-	-	-
BacHum F	-	-	-	-	-	-	246 (17%)	1654 (75%)	222 (4%)	-	-	-	-
BacHum R	268 (7%)	840 (12%)	1020 (9%)	1036 (8%)	-	-	260 (18%)	2720 (123%)	1239 (24%)	2640 (74%)	1036 (8%)	552 (30%)	537 (43%)
BacHum P	-	1 (0%)	7 (0%)	35 (0%)	-	-	246 (17%)	1858 (84%)	788 (15%)	2367 (67%)	-	-	-
HF183	-	-	-	-	-	-	252 (18%)	1662 (75%)	224 (4%)	-	-	-	-
BacCan 454F	-	-	-	-	-	-	-	-	-	-	-	-	-
DogBac DF475F	-	1 (0%)	-	4113 (33%)	-	-	-	-	1 (0%)	-	1 (0%)	-	-
Pig2Bac 41F	-	-	-	-	-	85 (4%)	-	-	-	-	-	-	-
Pig2Bac 163Rm	-	-	-	-	-	106 (4%)	-	-	-	-	-	-	-
Pig2Bac 113P	-	-	-	-	-	121 (5%)	-	-	-	-	-	-	-

Table 4.3: Source specificity for motif sequences towards faecal libraries. A positive result was assigned for faecal sources with at least 5% of all *Bacteroidales* sequences.

		Positive (≥5%)	Low level (2-5%)	Negative (≤2%)
Ruminant	BacCow CF128F	Alpaca, Cow, Sheep	Chicken	Dog, Horse, Pig, Possum, Human, Sewage, Duck, Pukeko, Swan
	BacR R	Alpaca, Cow, Sheep	Chicken, Duck	Dog, Horse, Pig, Possum, Human, Sewage, Pukeko, Swan
	BacR F	Alpaca, Cow, Pig, Sheep		Dog, Horse, Possum, Human, Sewage, Chicken, Duck, Pukeko, Swan
	BacR P	Alpaca, Cow, Sheep, Chicken	Dog, Horse, Sewage	Pig, Possum, Human, Duck, Pukeko, Swan
Human	BacH F	Human, Possum	Sewage	Alpaca, Cow, Sheep, Dog, Horse, Pig, Chicken, Duck, Pukeko, Swan
	BacH P-pC	Human, Possum	Sewage	Alpaca, Cow, Sheep, Dog, Horse, Pig, Chicken, Duck, Pukeko, Swan
	BacHum F	Human, Possum	Sewage	Alpaca, Cow, Sheep, Dog, Horse, Pig, Chicken, Duck, Pukeko, Swan
	HF183	Human, Possum	Sewage	Alpaca, Cow, Sheep, Dog, Horse, Pig, Chicken, Duck, Pukeko, Swan
	BacH P-oT	Chicken, Sewage, Human		Alpaca, Cow, Sheep, Dog, Horse, Pig, Possum, Duck, Pukeko, Swan
	BacHum P	Human, Chicken, Possum, Sewage		Alpaca, Cow, Sheep, Dog, Horse, Pig, Duck, Pukeko, Swan
Dog	DogBac DF475F	Dog		Alpaca, Cow, Sheep, Horse, Pig, Possum, Human, Sewage, Chicken, Duck, Pukeko, Swan
Pig	Pig2Bac 41F Pig2Bac 163Rm Pig2Bac 113P		Pig	Alpaca, Cow, Sheep, Dog, Horse, Possum, Human, Sewage, Chicken, Duck, Pukeko, Swan

Table 4.4: Sequence numbers and percentages for *Bacteroidales*-specific motifs screened against water samples. Percentages for non-source specific motifs are of total sequences, with percentages of source-specific motifs of total *Bacteroidales* sequences, determined by AllBac 296F sequences.

		NGS017	**NGS018**	**NGS126**	**NGS129**	**NGS130**	**NGS016**	**NGS125**	**NGS127**	**NGS128**	**NGS131**
	Total sequences	1488	11113	9523	4245	2098	1265	10796	8467	339	4952
Previously identified sources		Up to 100% Ruminant Contamination				10% Ruminant	Human	Human	Human Dog	Human	Human
Non-source specific	**AllBac 296F**	16 (1%)	204 (2%)	191 (2%)	161 (4%)	6 (0%)	118 (9%)	1249 (12%)	745 (9%)	11 (3%)	11 (0%)
	AllBac 412R	9 (1%)	112 (1%)	130 (1%)	77 (2%)	1 (0%)	151 (12%)	992 (9%)	583 (7%)	10 (3%)	5 (0%)
	AllBac 375P	31 (2%)	626 (6%)	371 (4%)	331 (8%)	165 (8%)	198 (16%)	1626 (15%)	1173 (14%)	20 (6%)	270 (5%)
	BacCow 305R	663 (45%)	1505 (14%)	1620 (17%)	918 (22%)	325 (15%)	269 (21%)	1316 (12%)	2366 (28%)	143 (42%)	670 (14%)
	BacCow 257P	22 (1%)	199 (2%)	192 (2%)	160 (4%)	6 (0%)	145 (11%)	1238 (11%)	739 (9%)	11 (3%)	11 (0%)
	BacHum R	1 (0%)	13 (0%)	33 (0%)	25 (1%)	-	88 (7%)	718 (7%)	115 (1%)	1 (0%)	4 (0%)
Ruminant	**BacCow CF128F**	2 (13%)	21 (10%)	12 (6%)	36 (22%)	-	1 (1%)	10 (1%)	-	-	-
Ruminant, Pig	**BacR F**	2 (13%)	16 (8%)	1 (1%)	29 (18%)	-	1 (1%)	-	-	-	-
Ruminant (Chicken, Duck)	**BacR R**	2 (13%)	18 (9%)	17 (9%)	20 (12%)	-	-	12 (1%)	-	-	-
Ruminant (Chicken)	**BacR P**	4 (25%)	26 (13%)	27 (14%)	75 (47%)	-	19 (16%)	17 (1%)	1 (0%)	-	2 (18%)
Human and Possum	**BacH F**	-	-	-	-	-	-	33 (3%)	22 (3%)	-	1 (9%)
	BacH P-pC	-	-	-	-	-	-	-	22 (3%)	-	1 (9%)
	BacHum F		-	-	-	-	-	38 (3%)	22 (3%)	-	1 (9%)
	HF183	-	-	-	-	-	-	39 (3%)	22 (3%)	-	1 (9%)
Human, Possum and Chicken	**BacHum P**	-	1 (0%)	2 (1%)	-	-	50 (42%)	602 (48%)	100 (13%)	1 (9%)	3 (27%)
Human and Chicken	**BacH P-oT**	-	1 (0%)	2 (1%)	-	-	80 (68%)	582 (47%)	85 (11%)	1 (9%)	2 (18%)
Dog marker	**DogBac DF475F**	-	-	-	-	-	-	-	24 (3%)	-	-
Pig markers	**Pig2Bac 41F**	-	2 (1%)	-	-	-	-	-	-	-	-
	Pig2Bac 163Rm	-	2 (1%)	-	-	-	-	-	-	-	-
	Pig2Bac 113P	-	2 (1%)	-	-	-	-	-	-	-	-

The cross-reactivity determined for each motif (Table 4.3) suggests that other faecal sources cannot be entirely ruled out. Of the four ruminant motif sequences analysed, two also match sequences in chicken faecal samples, with one also found in duck faeces. A third matches pig sequences as well as ruminants. This suggests that while we can be confident all three samples have ruminant contamination, there is also a possibility of contamination from these other three sources. The same can be applied to the human contaminated samples, with four markers also found in possum faecal samples, and two in chicken.

Table 4.5: Probable faecal contamination assignments for water samples.

		NGS018	**NGS126**	**NGS129**	**NGS016**	**NGS125**	**NGS127**
	Total sequences	11113	9523	4245	1265	10796	8467
Previously identified sources		Up to 100% Ruminant Contamination			Human		Human Dog
Non-source specific	**AllBac 296F**	204 (2%)	191 (2%)	161 (4%)	118 (9%)	1249 (12%)	745 (9%)
Ruminant	**BacCow CF128F**	Positive	Positive	Positive	-	-	-
Ruminant, Pig	**BacR F**	Positive	-	Positive	-	-	-
Ruminant (Chicken, Duck)	**BacR R**	Positive	Positive	Positive	-	-	-
Ruminant (Chicken)	**BacR P**	Positive	Positive	Positive	Positive	-	-
Human and Possum	**BacH F**	-	-	-	-	Positive	Positive
	BacH P-pC	-	-	-	-	-	Positive
	BacHum F	-	-	-	-	Positive	Positive
	HF183	-	-	-	-	Positive	Positive
Human, Possum and Chicken	**BacHum P**	-	-	-	Positive	Positive	Positive
Human and Chicken	**BacH P-oT**	-	-	-	Positive	Positive	Positive
Dog marker	**DogBac DF475F**	-	-	-	-	-	Positive
Pig markers	**Pig2Bac 41F**	2 (1%)	-	-	-	-	-
	Pig2Bac 163Rm	2 (1%)	-	-	-	-	-
	Pig2Bac 113P	2 (1%)	-	-	-	-	-

4.5 Discussion

NGS provides a valuable tool for studying microbial communities from a large range of environmental samples (Shokralla *et al.*, 2012). Most studies have focused on defining what bacteria are present, and what differences there are between microbial populations from different samples or environments. Very few studies have used NGS technologies

for a MST application, with most MST studies still using the wide range of source-specific assays which have been developed over the past decade.

4.5.1 Non-specific markers

As well as the three general *Bacteroidales* AllBac assay motif sequences, three additional motifs were found to be non-source specific. These were found in almost all faecal samples, with some samples showing higher numbers of sequence matches than the general AllBac markers. This trend was also seen in the water samples, particularly for the BacCow 305R motif sequence. The high proportion of this motif in all the water samples, as well as the sewage library, may suggest that this motif has targeted non-faecal origin *Bacteroidales*, matching sequences from environmental sources. Both water and sewage are known to contain environmental *Bacteroidales* bacteria.

4.5.2 Faecal source validation

A 5% threshold for identifying a positive source was evaluated against the faecal samples. This allowed all of the remaining source-specific motifs analysed to give a positive result for their sources, with the exception of the pig-specific motifs, which were found to represent 4-5%. Three motif sequences were not found in any source library, so were removed from further analysis.

Of the four ruminant motif sequences, none were found to be solely specific to ruminants. BacR F was also positive in pig samples, and BacR P was positive in chicken. Low-level cross-reactivity was also found in duck, dog, horse and sewage, with BacR P proving the least specific of the four markers.

The six human motif sequences were all positive in human, with two positive in sewage, and the remaining four motifs showing low-level detection. Interestingly, five motif sequences were also found in possum samples. This has also been seen in a recent PCR study where two human-specific markers had 100% false positive hits to brush-tailed possum faeces (Devane *et al.*, 2013). Two markers also showed positive results in chicken samples.

One of the two dog assays tested was found to be source specific to dog faeces, while all three markers from the pig assay tested were found only in pig faeces, suggesting that these markers are source-specific and suitable for assigning contamination sources using the motif-based search method proposed here, and may only require very low acceptance thresholds, as they have not been detected in any other faecal sample analysed.

One way of potentially increasing the specificity of this method would be to increase the number of sequences in each sample. The percentage of *Bacteroidales* sequences in each of the source libraries is lower than the suggested 30 – 40% previously suggested (Layton *et al.*, 2006). This may be due to sequence bias during the original amplicon PCR; however, taxonomy analysis for these samples implied that for the majority of sources, the number of sequences assigned to the *Bacteroidetes* phyla were around these values (Section 3.4.2.2), with the majority of these within the *Bacteroidales* order (data not shown). This may suggest that the AllBac assay may not be as universal to *Bacteroidales* as initially thought.

4.5.3 Water sample validation

Ten faecal contaminated water samples were screened using the successful primers (Tables 4.4 and 4.5) and suggested acceptance thresholds. Of these, four samples were removed from analysis due to not having enough sequences to allow a high level of confidence in the specificity of motif matches. NGS is not error-proof, and a single sequence match to a specific motif may be due to sequencing errors, rather than indicating the presence of a faecal source. Detection of a particular motif sequence or calculation of proportions may also be biased with low numbers of reads. In order to minimise this likelihood, we have suggested a sample to have a minimum of 1,000 sequences, and the minimum percentage of *Bacteroidales* sequences to be 2%. A higher minimum number would allow greater confidence in assigning contamination using these methods. For example, if a sample contained 3,000 sequences, it would allow a 95% probability of detecting a positive marker sequence if it is present at a prevalence of 0.1% of the total sample set (Midura and Bryant, 2001). However, this threshold level would result in an additional water sample also being removed (NGS016), which was able to be classified. If total sequence number was the only criteria, NGS131 would

be included, as it has more than 4,000 sequences, however, due to a very low percentage of *Bacteroidales* sequences, the motif screening results are unable to be confidently assigned to a contamination source.

The six water samples not eliminated due to low sequence numbers were all able to be assigned to a contamination source, using a 2% threshold for a positive result. No ruminant contaminated samples showed a positive result for human motif, and only one human contaminated sample contained a positive result for a ruminant marker (BacR P). This motif had been found to be positive for chicken sequences as well, which could suggest some chicken contamination in this water sample (NGS016). This water sample had not previously been tested for ruminant contamination, so it is also possible that it has a low level of ruminant contamination. NGS127 was the only water sample to contain sequences which matched the dog-specific assay sequence, which also correlates with previously found results.

4.5.4 Allowances for PCR primer mispairing

Variable conditions within PCR optimisation can often lead to mispairing of primers to their targets, with the annealing temperature an important factor (Dieffenbach *et al.*, 1993). The protocol used here screened for assay sequences with a perfect match, however, the number of mismatches allowed is able to be controlled by the user. Relaxing the mismatch threshold would allow for less specificity for each assay sequence, potentially resulting in a dramatic effect on sequence matches and their proportions to other source-specific assays. For example, allowing up to three mismatches in the assay screen for the swan sample resulted in a reasonable increase in all three AllBac assay sequences (Table 4.6), with sequences matching the probe marker more than five times greater. With no mismatches allowed, there were only matches to six assay markers, three of which were found in almost every source, and the other three matching less than 1% of all *Bacteroidales* sequences. However, by allowing mismatches, each of these three markers increase to representing more than 90% of the sequences, with an additional eight markers also found, showing high similarity to ruminant sources, with some similarity with human sources as well. There are also sequences assigned to the pig assay as well, which was only seen in pig faecal sources with no mismatches allowed. A relaxed threshold can also have large implications on

the water samples, which could add to the ability to assign a contamination source, or make it more complicated. Marker sequences do not increase proportionally, with some markers showing a much higher increase than others, such as the AllBac 375P assay sequence.

Table 4.6: Effects of mismatches allowed in primer sequence binding for the swan faecal source library.

	No mismatches	Up to 3 mismatches
AllBac 296F	1241 (11%)	1864 (16%)
AllBac 412R	1742 (15%)	1806 (15%)
AllBac 375P	1717 (15%)	9274 (79%)
BacCow CF128F	2 (0%)	1708 (92%)
BacCow 305R	1072 (86%)	6284 (337%)
BacCow 257P	1707 (138%)	1858 (100%)
BacR F	3 (0%)	1780 (95%)
BacR R	-	1025 (55%)
BacR P	3 (0%)	1844 (99%)
CF193	-	4 (0%)
BacH F	-	1 (0%)
BacH R	-	-
BacH P-pC	-	16 (1%)
BacH P-oT	-	33 (2%)
BacHum F	-	-
BacHum R	537 (43%)	1774 (95%)
BacHum P	-	530 (28%)
HF183	-	-
BacCan 454F	-	-
DogBac DF475F	-	-
Pig2Bac 41F	-	148 (8%)
Pig2Bac 163Rm	-	-
Pig2Bac 113P	-	70 (4%)

4.5.5 Effects of sequencing numbers

When using NGS amplicon sequencing techniques, samples are not uniformly sequenced to the same depth of coverage, resulting in a variation in the number of sequences for each sample. NGS018 and NGS125 – 131 were sequenced in the same sequencing run, with an additional eight samples. If sequencing had occurred evenly across the 16 samples, each sample would have had more than 8,800 sequence reads each. However, there was a large amount of variation, with samples ranging from 339 (NGS128) to 11,113 (NGS018). These differences in total number of sequences can be

seen to have an impact on the ability to detect assay marker sequences. NGS128 had very small numbers of AllBac assay sequences found, representing 3-6% of the total number of sequences, or 20 of the 339 sequences. This means that one sequence match to a source-specific marker is weighted at 9% of *Bacteroidales* sequences, which could result in a contamination source being assigned with very little actual sequence data to back it up. The low numbers can also mean that source-specific markers do not reach the recommended acceptance threshold, resulting in no contamination source being able to be assigned. This could lead to contamination issues not being addressed in the environment. It may also indicate that there is a low level of contamination.

The water samples can be divided into four categories based on read number requirements. The first contains the samples which had more than 3,000 total reads, and more than 100 AllBac 296F motif matches. Five of the ten water samples match these requirements, and can be confidently assigned to a contamination source using the source-specific motifs. NGS131 contains more than 3,000 total reads, but less than 100 AllBac 296F motif matches. This low proportion of *Bacteroidales* sequences suggests that there is only a very low level of faecal contamination. Three samples contain less than 3,000 reads and less than 100 AllBac 296F matches, suggesting that the total sequence numbers are too low to be able to assign contamination sources with confidence. NGS016 contained fewer than 3,000 reads, but more than 100 AllBac 296F matches, suggesting that it has a high proportion of fresh faecal contamination, and although contains less than 3,000 reads, can still be assigned to a contamination source with reasonable confidence.

4.5.6 DNA sequencing options

NGS sequencing technologies allow sequences to be generated in a number of ways. This study utilised amplicon sequencing, which requires a targeted DNA region to be amplified from the total DNA prior to sequencing. The benefits of this method are that it does not require a large amount of processing of the sequence data, as each sequence read is of the same target region. The Roche 454 platforms were chosen for this analysis because they allowed the longest sequencing read lengths. However, other platforms, such as the Illumina MiSeq and the Life Technologies Ion Torrent have also recently developed the capability to produce sequences of a similar read length. These platforms

achieve this through a paired-end approach, where target amplicons are sequenced from both ends, and matched together through overlaps in the middle regions. Other regions of the 16S rRNA gene can also be targeted, allowing other markers to be used.

An alternative to amplicon sequencing is to use whole genome sequencing. This method eliminates the need to amplify a target region by sequencing all the DNA present, reducing the potential of PCR bias. However, because of the limitations in sequencing length with the NGS platforms, it requires DNA to be fragmented into manageable sizes. After sequencing, each fragment needs to be positioned correctly in relation to the other fragments to produce the full genome. This is usually done through mapping sequence reads to a known alignment sequence. Whole genome sequencing for MST analysis would allow the sequences to be screened for a much larger range of assays, not limited to those which target the 16S rRNA gene. This could include pathogenic sequences, antibiotic resistance genes and virulence factors. This would allow water quality managers a more direct analysis of potential health risks. A disadvantage with this approach is that samples are often only sequenced to a low depth, which results in only the most dominant populations being observed (Zarraonaindia *et al.*, 2013). It is also much more computationally intense than amplicon sequencing, which requires more resources and time for analysis.

4.6 Conclusions

We have successfully provided a proof-of-concept study utilising NGS data and previously published *Bacteroidales* MST assays. The protocol outlined in this study provides a method that allows multiple contamination sources to be screened against any number of environmental samples, such a water sources. This removes the requirement of multiple PCR assays currently needed to assign more than one source of contamination. Four ruminant, six human, one dog and three pig marker sequences were found to provide source-specificity when a 2% threshold of the percentage of *Bacteroidales* sequences was used to determine a positive result. Six out of ten faecal contaminated water samples were correctly assigned to a contamination source using this protocol.

This compares well with the source-specific genera results from Chapter 3, where the contamination source of six of the water samples could be correctly identified (Table

3.9). Of these six, four provide the same result for both methods, with the contamination source of an additional two samples identified, which were not able to be using the methods outlined in this study. This is predominantly due to the different assessment criteria, with one study looking for specific genera, while the other only looks for specific *Bacteroidales* sequences. The source-specific genera identified in Chapter 3, or other NGS studies, could easily be included in the motif screens proposed in this study, with the number of motifs screened for only limited by the time available to screen each sample.

The computational motif-screening method proposed here is able to be scaled to different levels, depending on the requirements of the laboratory. There are an increasing number of software packages which are designed to work with NGS data, such as Geneious, which provide an easy interface for a range of people, and do not require an extensive knowledge of bioinformatics. The method can be applied to a range of areas, including MST, epidemiological and clinical studies, providing a backward compatibility method which can correlate health impacts with relevant markers.

Chapter Five
Summary and concluding remarks

5.1 Next generation sequencing

Next generation sequencing (NGS) is revolutionising environmental biology, making it possible to recover large amounts of sequence data from environmental samples. These techniques have been used in a variety of applications, including human (Andersson *et al.*, 2008; Costello *et al.*, 2009; Huse *et al.*, 2012) and animal (Lee *et al.*, 2011; Ley *et al.*, 2008a) microbiota studies, determination of bacterial biodiversity in a range of ecosystems (Lozupone and Knight, 2007; Sogin *et al.*, 2006; Tripathi *et al.*, 2012; Unno *et al.*, 2010) and diet analysis from faecal or gut contents (Boyer *et al.*, 2012; Deagle *et al.*, 2010). The implementation of barcoding techniques allows multiple samples to be sequenced simultaneously, increasing the high-throughput capabilities of NGS platforms. The number of samples included in a single sequencing run can influence the coverage of sequence reads obtained for each sample. In an ideal situation, each sample would have a similar number of reads, approximately the total number of sequences reads divided by the number of samples. The data presented in Chapter Three shows that this is not the case, with samples showing a large variation in sequence reads (Table 3.6). Samples with a low number of sequencing reads were much harder to assign to a contamination source (Chapter Four). This suggests that sequencing for multiple samples should be run at a much higher coverage than calculated, to ensure each sample has enough sequences for suitable analysis.

The cost of DNA sequencing has dropped dramatically over the past five years, as NGS technologies have largely taken over from the first generation Sanger sequencing. As the cost of sequencing continues to drop, the ability to produce large amounts of sequence data becomes more readily available to smaller facilities. Access to these low cost, high-throughput technologies can enable NGS to become incorporated into routine microbial diagnostic methods, such as clinical studies and environmental monitoring. For these methods to benefit the wider research community, standardised operating procedures should be put in place. This would remove the current necessity for laboratories to devise their own protocols, and allow multi-laboratory comparisons.

The Roche 454 pyrosequencing technology was used in this study, as it allowed for single direction amplicon-based sequencing which generated reads of approximately 500 nt in length. As other NGS platforms continue to be improved, many will have the potential to be as useful, if not more so, than the Roche 454 platforms. For example, the Illumina platforms have a much lower cost per base compared to the Roche 454 and generate a much larger output, but until recently, were unable to provide sequence read lengths greater than 150 bp. The use of paired-end reads enables this length to effectively be doubled, and recent advancements now enable 250 bp pair-end reads to be generated. The much larger data sets generated enable a much larger coverage compared with 454 sequencing, at a cost estimated to be approximately 50-fold lower (Bartram *et al.*, 2011).

5.2 Microbial source tracking

Traditional water quality monitoring has used the detection of indicator organisms such as *E. coli* or *enterococci*. These were the first targets of microbial source tracking using the creation of phenotyping or genotyping libraries of known sources to compare with water samples (Tallon *et al.* 2005). Limitations of this approach include the requirements for a very large library of isolates to represent the geographical and temporal variations in populations of microorganisms in each host species. It also became apparent that similarities in *E. coli* across multiple species, and changes in microbial populations over time, necessitate constant recreation of reference library of types.

These same issues may affect the approach described in Chapters Three and Four, where we looked at the differences in bacterial community compositions between different samples. Even though each sample was a composite of DNA extracted from five individual faecal samples, large variations were seen between individual samples from a single source (Section 3.4.2.5). The human microbiota has been found to have remarkable diversity at different sites within an individual, at the same site within an individual over time, and between different individuals (Parsley *et al.*, 2010). While the same bacteria may always be found, the proportions of different genera can change remarkably over time. This can result in differences in how samples cluster together when using diversity techniques, such as those described in Chapter Three, which may lead to different interpretations for similar data.

Bacterial diversity analysis of faecal samples used in this study shows that the host source is the largest influence in community diversity (Figures 3.7 and 3.8). Samples from the three ruminant sources all clustered tightly together, suggesting a very similar evolutionary history for the bacterial lineages found in each of these samples. These samples also tended to have the highest α-diversity (Figure 3.5), indicating a larger variation in species of bacteria is present within these sources.

Dog samples showed a large amount of β-diversity, particularly when a weighted UniFrac measure was applied, but showed the least amount of α-diversity. This demonstrates the large impact variation between individuals has on using these techniques to analyse diversity. The water samples analysed clustered together, with no indication of faecal contamination sources. Analysing only faecal-associated bacteria (Figures 3.9 and 3.10) did not provide any further information, suggesting that microbial diversity clustering techniques may not be suitable for analysis of faecal contamination in water sources. However, a number of faecal source-specific genera were found (Tables 3.8 and 3.9). Based on the relative presence of these genera, six of the ten faecal samples could be correctly attributed to a source of faecal contamination.

The 16S rRNA gene has been widely used for analysis of bacterial sequences due to its composition of both conserved and variable regions, which allow for species- or genera-level taxonomy identification (Chakravorty *et al.*, 2007). The V1-V3 hypervariable regions of the 16S rRNA gene were selected for analysis in this study, however, other regions have successfully been used for taxonomic identification in other metagenomic studies (e.g. (Degnan *et al.*, 2012; Flores *et al.*, 2012; McLellan *et al.*, 2010; Zhou *et al.*, 2011). Targeting different hypervariable regions has resulted in different taxonomic classifications for the same bacterial sequence (Kim *et al.*, 2011), with intrinsic biases towards certain taxa suggested to vary for different regions (Vilo and Dong, 2012). One way of avoiding these biases and potential taxonomic variations is to sequence the entire 16S rRNA gene. Using the full-length 1,500 nt sequence provides greater classification accuracies when compared to sequences in rRNA databases. Because of current length constraints in each of the NGS sequencing platforms, however, the full-length gene would have to be fragmented prior to sequencing, and the conserved nature of the gene makes it difficult to assemble these short reads into full-length 16S rRNA genes (Kim *et al.*, 2011).

A limitation with the use of partial 16S rRNA sequences is that they require PCR amplification to produce the desired target amplicons. The use of PCR introduces further bias, and does not produce a "true" representation of the bacteria communities actually present in the sample. It can be assumed that this bias is equal among all samples, provided they have been prepared using the same conditions. The data presented here were not always prepared in the same manner, with two different PCR protocols used. However, regardless of what protocol was used to prepare the sample, β-diversity analysis indicates that preparation protocol does not have a strong influence on diversity, with samples from the same source still clustering tightly together (Figure 3.7). The two samples from the same water sample, analysed using both PCR methods, were also found to cluster together when faecal bacteria were analysed (NGS018, Figure 3.10). As many of the faecal samples were prepared using DNA previously extracted and stored, the extraction protocols were not the same for all samples, which could have led to variability between samples. Ideally, studies utilising NGS amplicon-based sequencing should all be processed in the same way, over a similar time period, to remove the potential bias that may otherwise be introduced.

Taxonomic classification levels were not analysed to the level of species, due to the RDP database only classifying to the genus level. While some source-specific genera were identified (Table 3.7), taxonomic classification to the species level may be required to be able to identify more host-specific sequences. Alternative rRNA databases are available, and may provide better classification results. A marker which can differentiate between different ruminant sources, particularly between cow and sheep, is yet to be developed due to the high similarity in microbial composition.

The sequence motif search-based method outlined in Chapter Four provides an alternative method for analysing NGS data, utilising genetic marker sequences which have already proved successful in identifying contamination sources. Using this approach a 2% specificity threshold for the use of NGS data were determined, which allows reasonable specificity for the ruminant, human, and dog specific markers analysed. Four samples were removed from analysis due to limitations in the number of sequences. The remaining six samples were able to be assigned to the same contamination source as previously determined using a range of MST methods. Some level of non-specificity was detected with the selected motif sequences by screening them against the faecal libraries. Five of the six human-specific motif sequences also

matched sequences in the possum library, while two ruminant and two human motifs were found in the chicken library. One ruminant marker was also found in pig.

The development of source-specific molecular markers has reduced the need for the large databases of isolates, and allows microorganisms to be targeted regardless of whether they can be cultured in the laboratory (Su *et al.*, 2012). These molecular markers have also been found to be more geographically and temporally stable (Field and Samadpour, 2007). This has allowed for a greater understanding of what contamination is occurring within our waterways, but still does not show the complete picture of all potential risks. Because of the low numbers of these markers in environmental sources, these methods also require enrichment methods such as PCR to enable detection.

Compared with the analysis method in Chapter Three, the motif based approach may have the advantage of being less likely to be strongly influenced by an individual's variation in bacterial populations. It may also be more readily understood and used by water managers who are currently using the results of specific PCR based assays to understand sources of contamination.

The two strategies utilised with water samples (Chapters Three and Four) produced comparable results. While the same data sets have been used, it is reassuring that the different approaches to analysing them result in similar conclusions, and that these are consistent with previous MST analysis. Between the two strategies, eight of the ten water samples were assigned a contamination source, with four samples sharing the same result for both methods. Each method was also able to assign a source to an additional two samples, which were not identified by the other, suggesting there are advantages and limitations for both evaluation methods.

When analysing water samples, it is important to acknowledge the difference in the concentration of microorganisms compared with faecal samples, as faecal contamination will be diluted once in an aqueous environment. A low number of sequence reads from a sample may not be representative of the actual diversity in a sample. The expected level of diversity can be estimated through rarefaction curves, which can be used as a guideline for determining if enough sequence data has been gathered. If there are too few sequences, detection of a particular sequence or calculation of proportions may be biased. Therefore, a recommended minimum number

of sequences required for a sample to be analysed would be beneficial when using these methods. For example, if a sample contained 3,000 sequences, it would allow a 95% probability of detecting a positive marker sequence if it is present at a prevalence of 0.1% of the total sample set (Midura and Bryant, 2001). This threshold would result in four of the ten water samples being removed from analysis, as well as one of each of the two water samples which were included twice.

5.3 Other applications and future directions

The methods outlined in this thesis could easily be applied to microbial diversity analysis in other environmental samples. An example of this is aquifer microcosms, where DNA extracted from sediment cores can be analysed for microbial communities. Two aquifer microcosm samples plus a groundwater sample used to flush the microcosms were included in the sequencing runs used for the analysis in the study outlined here. While the sequencing data were not included in this study, the same analysis protocols outlined in Chapter Three were applied to evaluate the microbial community of these samples. Information on the bacteria found in the sediments of aquifers can be used by water managers to ensure there is no contamination entering the aquifers, which could potentially lead to human health issues.

The best-case scenario for moving forward with environmental monitoring techniques would be to sequence everything, by extracting DNA directly from the sample, with no enrichment required. This would allow for a much broader analysis and understanding of what is happening, including direct detection of pathogens, and will have the potential for large scale automation. It would also allow for environmental changes to be monitored, which may provide additional information about what is happening in our environment.

References

Abu Al-Soud, W. and Rådström, P. (1998). Capacity of nine thermostable DNA polymerases to mediate DNA amplification in the presence of PCR-inhibiting samples. *Applied and Environmental Microbiology* **64**, 3748-3753.

Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermod, J. J., Mayer, P. and Kawashima, E. (2000). Solid phase DNA amplification: Characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Research* **28**, e87.

Ahmed, W., Stewart, J., Gardner, T., Powell, D., Brooks, P., Sullivan, D. and Tindale, N. (2007). Sourcing faecal pollution: A combination of library-dependent and library-independent methods to identify human faecal pollution in non-sewered catchments. *Water Research* **41**, 3771-3779.

Ahmed, W., Stewart, J., Powell, D. and Gardner, T. (2008). Evaluation of the host-specificity and prevalence of Enterococci Surface Protein (*esp*) marker in sewage and its application for sourcing human fecal pollution. *Journal of Environmental Quality* **37**, 1583-1588.

Amann, R. I., Ludwig, W. and Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews* **59**, 143-169.

Amend, A. S., Seifert, K. A. and Bruns, T. D. (2010). Quantifying microbial communities with 454 pyrosequencing: Does read abundance count? *Molecular Ecology* **19**, 5555-5565.

Anderson, M. A., Whitlock, J. E. and Harwood, V. J. (2006). Diversity and distribution of *Escherichia coli* genotypes and antibiotic resistance phenotypes in feces of humans, cattle, and horses. *Applied and Environmental Microbiology* **72**, 6914-6922.

Andersson, A. F., Lindberg, M., Jakobsson, H., Bäckhed, F., Nyrén, P. and Engstrand, L. (2008). Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS ONE* **3**, e2836.

Archer, J., Weber, J., Henry, K., Winner, D., Gibson, R., Lee, L., Paxinos, E., Arts, E. J., *et al.* (2012). Use of four next-generation sequencing platforms to determine HIV-1 coreceptor tropism. *PLoS ONE* **7**, e49602.

Baker, G. C., Smith, J. J. and Cowan, D. A. (2003). Review and re-analysis of domain-specific 16S primers. *Journal of Microbiological Methods* **55**, 541-555.

Bartram, A. K., Lynch, M. D. J., Stearns, J. C., Moreno-Hagelsieb, G. and Neufeld, J. D. (2011). Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. *Applied and Environmental Microbiology* **77**, 3846-3852.

Beloshapka, A. N., Dowd, S. E., Suchodolski, J. S., Steiner, J. M., Duclos, L. and Swanson, K. S. (2013). Fecal microbial communities of healthy adult dogs fed raw meat-based diets with or without inulin or yeast cell wall extracts as assessed by 454 pyrosequencing. *FEMS Microbiology Ecology* **84**, 532-541.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59.

Bernhard, A. E. and Field, K. G. (2000a). Identification of nonpoint sources of fecal pollution in coastal waters by using host-specific 16S ribosomal DNA genetic

markers from fecal anaerobes. *Applied and Environmental Microbiology* **66**, 1587-1594.

Bernhard, A. E. and Field, K. G. (2000b). A PCR assay to discriminate human and ruminant feces on the basis of host differences in *Bacteroides-Prevotella* genes encoding 16S rRNA. *Applied and Environmental Microbiology* **66**, 4571-4574.

Binladen, J., Gilbert, M. T. P., Bollback, J. P., Panitz, F., Bendixen, C., Nielsen, R. and Willerslev, E. (2007). The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE* **2**, e197.

Bodrossy, L. and Sessitsch, A. (2004). Oligonucleotide microarrays in microbial diagnostics. *Current Opinion in Microbiology* **7**, 245-254.

Boehm, A. B. (2007). Enterococci concentrations in diverse coastal environments exhibit extreme variability. *Environmental Science and Technology* **41**, 8227-8232.

Boehm, A. B., Yamahara, K. M., Love, D. C., Peterson, B. M., McNeill, K. and Nelson, K. L. (2009). Covariation and photoinactivation of traditional and novel indicator organisms and human viruses at a sewage-impacted marine beach. *Environmental Science and Technology* **43**, 8046-8052.

Boers, S. A., van der Reijden, W. A. and Jansen, R. (2012). High-throughput multilocus sequence typing: Bringing molecular typing to the next level. *PLoS ONE* **7**, e39630.

Bonjoch, X., Lucena, F. and Blanch, A. R. (2009). The persistence of *bifidobacteria* populations in a river measured by molecular and culture techniques. *Journal of Applied Microbiology* **107**, 1178-1185.

Booth, A. M., Hagedorn, C., Graves, A. K., Hagedorn, S. C. and Mentz, K. H. (2003). Sources of fecal pollution in Virginia's Blackwater River. *Journal of Environmental Engineering* **129**, 547-552.

Bowers, J., Mitchell, J., Beer, E., Buzby, P. R., Causey, M., Efcavitch, J. W., Jarosz, M., Krzymanska-Olejnik, E., *et al.* (2009). Virtual terminator nucleotides for next-generation DNA sequencing. *Nature Methods* **6**, 593-595.

Bowers, R. M., McLetchie, S., Knight, R. and Fierer, N. (2011). Spatial variability in airborne bacterial communities across land-use types and their relationship to the bacterial communities of potential source environments. *ISME Journal* **5**, 601-612.

Boyer, S., Brown, S. D. J., Collins, R. A., Cruickshank, R. H., Lefort, M. C., Malumbres-Olarte, J. and Wratten, S. D. (2012). Sliding window analyses for optimal selection of mini-barcodes, and application to 454-pyrosequencing for specimen identification from degraded DNA. *PLoS ONE* **7**, e38215.

Bruce, K. D. (1997). Analysis of mer gene subclasses within bacterial communities in soils and sediments resolved by fluorescent-PCR-restriction fragment length polymorphism profiling. *Applied and Environmental Microbiology* **63**, 4914-4919.

Buchan, A., Alber, M. and Hodson, R. E. (2001). Strain-specific differentiation of environmental *Escherichia coli* isolates via denaturing gradient gel electrophoresis (DGGE) analysis of the 16S-23S intergenic spacer region. *FEMS Microbiology Ecology* **35**, 313-321.

Byappanahalli, M. N., Przybyla-Kelly, K., Shively, D. A. and Whitman, R. L. (2008). Environmental occurrence of the enterococcal surface protein (*esp*) gene is an unreliable indicator of human fecal contamination. *Environmental Science and Technology* **42**, 8014-8020.

Bystrykh, L. V. (2012). Generalized DNA barcode design based on Hamming codes. *PLoS ONE* **7**, e36852.

Cabelli, V. J. (1983). Health effects criteria for marine recreational waters. *Health Effects Research Laboratory, Office of Research and Development, US Environmental Protection Agency* **EPA 600/1-80-031,** August.

Caldwell, J., Payment, P. and Villemur, R. (2011). Mitochondrial DNA as source tracking markers. In *Microbial source tracking: Methods, applications and case studies*, pp. 229-250. Edited by C. Hagedorn, V. J. Harwood and A. R. Blanch. New York, USA: Springer Science+Business Media.

Cao, Y., Van De Werfhorst, L. C., Sercu, B., Murray, J. L. S. and Holden, P. A. (2011). Application of an integrated community analysis approach for microbial source tracking in a coastal creek. *Environmental Science and Technology* **45**, 7195-7201.

Caporaso, J. G., Bittinger, K., Bushman, F. D., Desantis, T. Z., Andersen, G. L. and Knight, R. (2009). PyNAST: A flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**, 266-267.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Pẽa, A. G., *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**, 335-336.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S. M., Betley, J., *et al.* (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME Journal* **6**, 1621-1624.

Cardenas, E. and Tiedje, J. M. (2008). New tools for discovering and characterizing microbial diversity. *Current Opinion in Biotechnology* **19**, 544-549.

Carroll, I. M., Ringel-Kulka, T., Siddle, J. P., Klaenhammer, T. R. and Ringel, Y. (2012). Characterization of the fecal microbiota using high-throughput sequencing reveals a stable microbial community during storage. *PLoS ONE* **7**, e46953.

Casarez, E. A., Pillai, S. D. and Di Giovanni, G. D. (2007a). Genotype diversity of *Escherichia coli* isolates in natural waters determined by PFGE and ERIC-PCR. *Water Research* **41**, 3643-3648.

Casarez, E. A., Pillai, S. D., Mott, J. B., Vargas, M., Dean, K. E. and Di Giovanni, G. D. (2007b). Direct comparison of four bacterial source tracking methods and use of composite data sets. *Journal of Applied Microbiology* **103**, 350-364.

Case, R. J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W. F. and Kjelleberg, S. (2007). Use of 16S rRNA and *rpoB* genes as molecular markers for microbial ecology studies. *Applied and Environmental Microbiology* **73**, 278-288.

Chakravorty, S., Helb, D., Burday, M., Connell, N. and Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of Microbiological Methods* **69**, 330-339.

Chen, Z., Pavelic, P., Dillon, P. and Naidu, R. (2002). Determination of caffeine as a tracer of sewage effluent in natural waters by on-line solid-phase extraction and liquid chromatography with diode-array detection. *Water Research* **36**, 4830-4838.

Cheung, W. H. S., Chang, K. C. K., Hung, R. P. S. and Kleevens, J. W. L. (1990). Health effects of beach water pollution in Hong Kong. *Epidemiology and Infection* **105**, 139-162.

Choi, S., Chu, W., Brown, J., Becker, S. J., Harwood, V. J. and Jiang, S. C. (2003). Application of enterococci antibiotic resistance patterns for contamination source identification at Huntington Beach, California. *Marine Pollution Bulletin* **46**, 748-755.

Chun, J., Kim, K. Y., Lee, J. H. and Choi, Y. (2010). The analysis of oral microbial communities of wild-type and toll-like receptor 2-deficient mice using a 454 GS FLX Titanium pyrosequencer. *BMC Microbiology* **10**, 101.

Claesson, M. J., O'Sullivan, O., Wang, Q., Nikkilä, J., Marchesi, J. R., Smidt, H., de Vos, W. M., Ross, R. P. and O'Toole, P. W. (2009). Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PLoS ONE* **4**, e6669.

Claesson, M. J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J. R., Ross, R. P. and O'Toole, P. W. (2010). Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Research* **38**, e200.

Clarke, J., Wu, H. C., Jayasinghe, L., Patel, A., Reid, S. and Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology* **4**, 265-270.

Clarridge III, J. E. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical Microbiology Reviews* **17**, 840-862.

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., *et al.* (2009). The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Research* **37**, D141-D145.

Collins, F. S., Lander, E. S., Rogers, J. and Waterson, R. H. (2004). Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945.

Converse, R. R., Blackwood, A. D., Kirs, M., Griffith, J. F. and Noble, R. T. (2009). Rapid QPCR-based assay for fecal *Bacteroides* spp. as a tool for assessing fecal contamination in recreational waters. *Water Research* **43**, 4828-4837.

Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I. and Knight, R. (2009). Bacterial community variation in human body habitats across space and time. *Science* **326**, 1694-1697.

Cox-Foster, D. L., Conlan, S., Holmes, E. C., Palacios, G., Evans, J. D., Moran, N. A., Quan, P. L., Briese, T., *et al.* (2007). A metagenomic survey of microbes in honey bee colony collapse disorder. *Science* **318**, 283-287.

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J. B., Massart, S., Collini, S., Pieraccini, G. and Lionetti, P. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 14691-14696.

Deagle, B. E., Chiaradia, A., McInnes, J. and Jarman, S. N. (2010). Pyrosequencing faecal DNA to determine diet of little penguins: Is what goes in what comes out? *Conservation Genetics* **11**, 2039-2048.

Degnan, P. H., Pusey, A. E., Lonsdorf, E. V., Goodall, J., Wroblewski, E. E., Wilson, M. L., Rudicell, R. S., Hahn, B. H. and Ochman, H. (2012). Factors associated with the diversification of the gut microbial communities within chimpanzees from Gombe National Park. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 13034-13039.

DeSantis, T. Z., Brodie, E. L., Moberg, J. P., Zubieta, I. X., Piceno, Y. M. and Andersen, G. L. (2007). High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. *Microbial Ecology* **53**, 371-383.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., *et al.* (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology* **72**, 5069-5072.

Dethlefsen, L., Huse, S., Sogin, M. L. and Relman, D. A. (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16s rRNA sequencing. *PLoS Biology* **6**, 2383-2400.

Devane, M., Robson, B., Nourozi, F., Wood, D. and Gilpin, B. J. (2013). Distinguishing human and possum faeces using PCR markers. *Journal of Water and Health* **In press**, doi: 10.2166/wh.2013.2122.

Dick, L. K., Bernhard, A. E., Brodeur, T. J., Santo Domingo, J. W., Simpson, J. M., Walters, S. P. and Field, K. G. (2005a). Host distributions of uncultivated fecal *Bacteroidales* bacteria reveal genetic markers for fecal source identification. *Applied and Environmental Microbiology* **71**, 3184-3191.

Dick, L. K. and Field, K. G. (2004). Rapid estimation of numbers of fecal *Bacteroidetes* by use of a quantitative PCR assay for 16S rRNA genes. *Applied and Environmental Microbiology* **70**, 5695-5697.

Dick, L. K., Simonich, M. T. and Field, K. G. (2005b). Microplate subtractive hybridization to enrich for *Bacteroidales* genetic markers for fecal source identification. *Applied and Environmental Microbiology* **71**, 3179-3183.

Dick, L. K., Stelzer, E. A., Bertke, E. E., Fong, D. L. and Stoeckel, D. M. (2010). Relative decay of *Bacteroidales* microbial source tracking markers and cultivated *Escherichia coli* in freshwater microcosms. *Applied and Environmental Microbiology* **76**, 3255-3262.

Dickerson Jr, J. W., Hagedorn, C. and Hassall, A. (2007). Detection and remediation of human-origin pollution at two public beaches in Virginia using multiple source tracking methods. *Water Research* **41**, 3758-3770.

Dieffenbach, C. W., Lowe, T. M. J. and Dveksler, G. S. (1993). General concepts for PCR primer design. *Genome Research* **3**, S30-S37.

Dohm, J. C., Lottaz, C., Borodina, T. and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research* **36**, e105.

Dombek, P. E., Johnson, L. K., Zimmerley, S. T. and Sadowsky, M. J. (2000). Use of repetitive DNA sequences and the PCR to differentiate *Escherichia coli* isolates from human and animal sources. *Applied and Environmental Microbiology* **66**, 2572-2577.

Dorai-Raj, S., Grady, J. O. and Colleran, E. (2009). Specificity and sensitivity evaluation of novel and existing *Bacteroidales* and *Bifidobacteria*-specific PCR assays on feces and sewage samples and their application for microbial source tracking in Ireland. *Water Research* **43**, 4980-4988.

Dowd, S. E., Callaway, T. R., Wolcott, R. D., Sun, Y., McKeehan, T., Hagevoort, R. G. and Edrington, T. S. (2008). Evaluation of the bacterial diversity in the feces of cattle using 16S rDNA bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP). *BMC Microbiology* **8**, e125.

Dressman, D., Yan, H., Traverso, G., Kinzler, K. W. and Vogelstein, B. (2003). Transforming single DNA molecules into fluorescent magnetic particles for

detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 8817-8822.

Dubinsky, E. A., Esmaili, L., Hulls, J. R., Cao, Y., Griffith, J. F. and Andersen, G. L. (2012). Application of phylogenetic microarray analysis to discriminate sources of fecal pollution. *Environmental Science and Technology* **46**, 4340-4347.

Duran, M., Haznedaroğlu, B. Z. and Zitomer, D. H. (2006). Microbial source tracking using host specific FAME profiles of fecal coliforms. *Water Research* **40**, 67-74.

Durso, L. M., Wells, J. E., Harhay, G. P., Rice, W. C., Kuehn, L., Bono, J. L., Shackelford, S., Wheeler, T. and Smith, T. P. L. (2012). Comparison of bacterial communities in faeces of beef cattle fed diets containing corn and wet distillers' grain with solubles. *Letters in Applied Microbiology* **55**, 109-114.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461.

Edwards, R. A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D. M., Saar, M. O., Alexander, S., *et al.* (2006). Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7**, e57.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., *et al.* (2009). Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138.

Ekblom, R. and Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* **107**, 1-15.

Emrich, C. A., Tian, H., Medintz, I. L. and Mathies, R. A. (2002). Microfabricated 384-lane capillary array electrophoresis bioanalyzer for ultrahigh-throughput genetic analysis. *Analytical Chemistry* **74**, 5076-5083.

Engberg, J., On, S. L. W., Harrington, C. S. and Gerner-Smidt, P. (2000). Prevalence of *Campylobacter*, *Arcobacter*, *Helicobacter*, and *Sutterella* spp. in human fecal samples as estimated by a reevaluation of isolation methods for Campylobacters. *Journal of Clinical Microbiology* **38**, 286-291.

Erlich, Y., Mitra, P. P., delaBastide, M., McCombie, W. R. and Hannon, G. J. (2008). Alta-Cyclic: A self-optimizing base caller for next-generation sequencing. *Nature Methods* **5**, 679-682.

Farrelly, V., Rainey, F. A. and Stackebrandt, E. (1995). Effect of genome size and *rrn* gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. *Applied and Environmental Microbiology* **61**, 2798-2801.

Fedurco, M., Romieu, A., Williams, S., Lawrence, I. and Turcatti, G. (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Research* **34**.

Field, K. G., Chern, E. C., Dick, L. K., Fuhrman, J., Griffith, J., Holden, P. A., LaMontagne, M. G., Le, J., *et al.* (2003). A comparative study of culture-independent, library-independent genotypic methods of fecal source tracking. *Journal of Water and Health* **1**, 181-194.

Field, K. G. and Samadpour, M. (2007). Fecal source tracking, the indicator paradigm, and managing water quality. *Water Research* **41**, 3517-3538.

Fierer, N., Breitbart, M., Nulton, J., Salamon, P., Lozupone, C., Jones, R., Robeson, M., Edwards, R. A., *et al.* (2007). Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Applied and Environmental Microbiology* **73**, 7059-7066.

Fierer, N., Hamady, M., Lauber, C. L. and Knight, R. (2008). The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 17994-17999.

Finegold, S. M., Vaisanen, M. L., Molitoris, D. R., Tomzynski, T. J., Song, Y., Liu, C., Collins, M. D. and Lawson, P. A. (2003). *Cetobacterium somerae* sp. nov. from human feces and emended description of the genus *Cetobacterium*. *Systematic and Applied Microbiology* **26**, 177-181.

Flores, G. E., Henley, J. B. and Fierer, N. (2012). A direct PCR approach to accelerate analyses of human-associated microbial communities. *PLoS ONE* **7**, e44563.

Fogarty, L. R. and Voytek, M. A. (2005). Comparison of *Bacteroides-Prevotella* 16S rRNA genetic markers for fecal samples from different animal species. *Applied and Environmental Microbiology* **71**, 5999-6007.

Fong, T. T., Griffin, D. W. and Lipp, E. K. (2005). Molecular assays for targeting human and bovine enteric viruses in coastal waters and their application for library-independent source tracking. *Applied and Environmental Microbiology* **71**, 2070-2078.

Foster, G., Ross, H. M., Naylor, R. D., Collins, M. D., Pascual Ramos, C., Fernandez Garayzabal, F. and Reid, R. J. (1995). *Cetobacterium ceti* gen. nov., sp. nov., a new Gram-negative obligate anaerobe from sea mammals. *Letters in Applied Microbiology* **21**, 202-206.

Fremaux, B., Boa, T. and Yost, C. K. (2010). Quantitative real-time PCR assays for sensitive detection of canada goose-specific fecal pollution in water sources. *Applied and Environmental Microbiology* **76**, 4886-4889.

Fremaux, B., Gritzfeld, J., Boa, T. and Yost, C. K. (2009). Evaluation of host-specific *Bacteroidales* 16S rRNA gene markers as a complementary tool for detecting fecal pollution in a prairie watershed. *Water Research* **43**, 4838-4849.

Genthner, F. J., James, J. B., Yates, D. F. and Friedman, S. D. (2005). Use of composite data sets for source-tracking enterococci in the water column and shoreline interstitial waters on Pensacola Beach, Florida. *Marine Pollution Bulletin* **50**, 724-732.

Gilpin, B., James, T., Nourozi, F., Saunders, D., Scholes, P. and Savill, M. (2003). The use of chemical and molecular microbial indicators for faecal source identification. *Water Science and Technology* **47**, 39-43.

Gilpin, B. J., Gregor, J. E. and Savill, M. G. (2002). Identification of the source of faecal pollution in contaminated rivers. *Water Science and Technology* **46**, 9-15.

Glassmeyer, S. T., Furlong, E. T., Kolpin, D. W., Cahill, J. D., Zaugg, S. D., Werner, S. L., Meyer, M. T. and Kryak, D. D. (2005). Transport of chemical and microbial compounds from known wastewater discharges: Potential for use as indicators of human fecal contamination. *Environmental Science and Technology* **39**, 5157-5169.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* **11**, 759-769.

Gonzalez, A. and Knight, R. (2012). Advancing analytical algorithms and pipelines for billions of microbial sequences. *Current Opinion in Biotechnology* **23**, 64-71.

Gordon, D. M. (2001). Geographical structure and host specificity in bacteria and the implications for tracing the source of coliform contamination. *Microbiology* **147**, 1079-1085.

Gordon, D. M., Bauer, S. and Johnson, J. R. (2002). The genetic structure of *Escherichia coli* populations in primary and secondary habitats. *Microbiology* **148**, 1513-1522.

Graczyk, T. K., Fayer, R., Trout, J. M., Lewis, E. J., Farley, C. A., Sulaiman, I. and Lal, A. A. (1998). *Giardia* sp. cysts and infectious *Cryptosporidium parvum* oocysts in the feces of migratory Canada geese (*Branta canadensis*). *Applied and Environmental Microbiology* **64**, 2736-2738.

Graves, A. K., Hagedorn, C., Brooks, A., Hagedorn, R. L. and Martin, E. (2007). Microbial source tracking in a rural watershed dominated by cattle. *Water Research* **41**, 3729-3739.

Green, H. C., Dick, L. K., Gilpin, B., Samadpour, M. and Field, K. G. (2012). Genetic markers for rapid PCR-based identification of gull, Canada goose, duck, and chicken fecal contamination in water. *Applied and Environmental Microbiology* **78**, 503-510.

Green, R. E., Krause, J., Ptak, S. E., Briggs, A. W., Ronan, M. T., Simons, J. F., Du, L., Egholm, M., *et al.* (2006). Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**, 330-336.

Greetham, H. L., Collins, M. D., Gibson, G. R., Giffard, C., Falsen, E. and Lawson, P. A. (2004). *Sutterella stercoricanis* sp. nov., isolated from canine faeces. *International Journal of Systematic and Evolutionary Microbiology* **54**, 1581-1584.

Gregor, J., Garrett, N., Gilpin, B., Randall, C. and Saunders, D. (2002). Use of classification and regression tree (CART) analysis with chemical faecal indicators to determine sources of contamination. *New Zealand Journal of Marine and Freshwater Research* **36**, 387-398.

Griffith, J. F., Weisberg, S. B. and McGee, C. D. (2003). Evaluation of microbial source tracking methods using mixed fecal sources in aqueous test samples. *Journal of Water and Health* **1**, 141-151.

Guan, S., Xu, R., Chen, S., Odumeru, J. and Gyles, C. (2002). Development of a procedure for discriminating among *Escherichia coli* isolates from animal and human sources. *Applied and Environmental Microbiology* **68**, 2690-2698.

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., Ciulla, D., Tabbaa, D., *et al.* (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research* **21**, 494-504.

Hagedorn, C., Crozier, J. B., Mentz, K. A., Booth, A. M., Graves, A. K., Nelson, N. J. and Reneau Jr, R. B. (2003). Carbon source utilization profiles as a method to identify sources of faecal pollution in water. *Journal of Applied Microbiology* **94**, 792-799.

Hagedorn, C., Harwood, V. J. and Blanch, A. R. (2011a). Overview. In *Microbial source tracking: Methods, applications, and case studies*, pp. 1-6. Edited by C. Hagedorn, V. J. Harwood and A. R. Blanch. New York, USA: Springer Science+Business Media.

Hagedorn, C. and Liang, X. (2011). Current and future trends in fecal source tracking and deployment in the Lake Taihu Region of China. *Physics and Chemistry of the Earth* **36**, 352-359.

Hagedorn, C., Robinson, S. L., Filtz, J. R., Grubbs, S. M., Angier, T. A. and Reneau Jr, R. B. (1999). Determining sources of fecal pollution in a rural Virginia watershed with antibiotic resistance patterns in fecal streptococci. *Applied and Environmental Microbiology* **65**, 5522-5531.

Hagedorn, C. and Weisberg, S. B. (2009). Chemical-based fecal source tracking methods: Current status and guidelines for evaluation. *Reviews in Environmental Science and Biotechnology* **8**, 275-287.

Hagedorn, C., Weisberg, S. B., Blanch, A. R. and Harwood, V. J. (2011b). Chemical-based fecal source tracking methods. In *Microbial source tracking: Methods, applications, and case studies*, pp. 189-206. Edited by C. Hagedorn, V. J. Harwood and A. R. Blanch. New York, USA: Springer Science+Business Media.

Hamady, M., Walker, J. J., Harris, J. K., Gold, N. J. and Knight, R. (2008). Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nature Methods* **5**, 235-237.

Hamilton, M. J., Yan, T. and Sadowsky, M. J. (2006). Development of goose- and duck-specific DNA markers to determine sources of *Escherichia coli* in waterways. *Applied and Environmental Microbiology* **72**, 4012-4019.

Handelsman, J. (2004). Metagenomics: Application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews* **68**, 669-685.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chemistry and Biology* **5**, R245-R249.

Handl, S., Dowd, S. E., Garcia-Mazcorro, J. F., Steiner, J. M. and Suchodolski, J. S. (2011). Massive parallel 16S rRNA gene pyrosequencing reveals highly diverse fecal bacterial and fungal communities in healthy dogs and cats. *FEMS Microbiology Ecology* **76**, 301-310.

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., Schork, N. J., Murray, S. S., *et al.* (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology* **10**, R32.

Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., *et al.* (2008). Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106-109.

Harwood, V. J. (2007). Assumptions and limitations associated with microbial source tracking methods. In *Microbial source tracking*, pp. 33-64. Edited by J. W. Santo Domingo and M. J. Sadowsky. Washington, DC, USA: ASM Press.

Harwood, V. J., Delahoya, N. C., Ulrich, R. M., Kramer, M. F., Whitlock, J. E., Garey, J. R. and Lim, D. V. (2004). Molecular confirmation of *Enterococcus faecalis* and *E. faecium* from clinical, faecal and environmental sources. *Letters in Applied Microbiology* **38**, 476-482.

Harwood, V. J. and Stoeckel, D. M. (2011). Performance criteria. In *Microbial source tracking: Methods, applications, and case studies*, pp. 7-30. Edited by C. Hagedorn, A. R. Blanch and V. J. Harwood. New York, USA: Springer Science+Business Media.

Harwood, V. J., Whitlock, J. and Withington, V. (2000). Classification of antibiotic resistance patterns of indicator bacteria by discriminant analysis: Use in predicting the source of fecal contamination in subtropical waters. *Applied and Environmental Microbiology* **66**, 3698-3704.

Harwood, V. J., Wiggins, B., Hagedorn, C., Ellender, R. D., Gooch, J., Kern, J., Samadpour, M., Chapman, A. C., *et al.* (2003). Phenotypic library-based microbial source tracking methods: Efficacy in the California collaborative study. *Journal of Water and Health* **1**, 153-166.

Haznedaroğlu, B. Z., Zitomer, D. H., Hughes-Strange, G. B. and Duran, M. (2005). Whole-cell fatty acid composition of total coliforms to predict sources of fecal contamination. *Journal of Environmental Engineering* **131**, 1426-1432.

Hert, D. G., Fredlake, C. P. and Barron, A. E. (2008). Advantages and limitations of next-generation sequencing technologies: A Comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis* **29**, 4618-4626.

Heylen, K., Lebbe, L. and de Vos, P. (2008). *Acidovorax caeni* sp. nov., a denitrifying species with genetically diverse isolates from activated sludge. *International Journal of Systematic and Evolutionary Microbiology* **58**, 73-77.

Hill, J. E., Ursla Fernando, W. M., Zello, G. A., Tyler, R. T., Dahl, W. J. and Van Kessel, A. G. (2010). Improvement of the representation of Bifidobacteria in fecal microbiota metagenomic libraries by application of the *cpn60* universal primer cocktail. *Applied and Environmental Microbiology* **76**, 4550-4552.

Hilton, M. J. and Thomas, K. V. (2003). Determination of selected human pharmaceutical compounds in effluent and surface water samples by high-performance liquid chromatography-electrospray tandem mass spectrometry. *Journal of Chromatography A* **1015**, 129-141.

Hoffmann, C., Minkah, N., Leipzig, J., Wang, G., Arens, M. Q., Tebas, P. and Bushman, F. D. (2007). DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Research* **35**, e91.

Holdeman, L. V., Good, I. J. and Moore, W. E. C. (1976). Human fecal flora: Variation in bacterial composition within individuals and a possible effect of emotional stress. *Applied and Environmental Microbiology* **31**, 359-375.

Housby, J. N. and Southern, E. M. (1998). Fidelity of DNA ligation: A novel experimental approach based on the polymerisation of libraries of oligonucleotides. *Nucleic Acids Research* **26**, 4259-4266.

Howorka, S., Cheley, S. and Bayley, H. (2001). Sequence-specific detection of individual DNA strands using engineered nanopores. *Nature Biotechnology* **19**, 636-639.

Huber, J. A., Mark Welch, D. B., Morrison, H. G., Huse, S. M., Neal, P. R., Butterfield, D. A. and Sogin, M. L. (2007). Microbial population structures in the deep marine biosphere. *Science* **318**, 97-100.

Hugenholtz, P. and Tyson, G. W. (2008). Metagenomics. *Nature* **455**, 481-483.

Hughes, J. B., Hellmann, J. J., Ricketts, T. H. and Bohannan, B. J. M. (2001). Counting the uncountable: Statistical approaches to estimating microbial diversity. *Applied and Environmental Microbiology* **67**, 4399-4406.

Hulcr, J., Latimer, A. M., Henley, J. B., Rountree, N. R., Fierer, N., Lucky, A., Lowman, M. D. and Dunn, R. R. (2012). A jungle in there: Bacteria in belly buttons are highly diverse, but predictable. *PLoS ONE* **7**, e47713.

Hundesa, A., Bofill-Mas, S., Maluquer de Motes, C., Rodriguez-Manzano, J., Bach, A., Casas, M. and Girones, R. (2010). Development of a quantitative PCR assay for the quantitation of bovine polyomavirus as a microbial source-tracking tool. *Journal of Virological Methods* **163**, 385-389.

Huse, S. M., Welch, D. M., Morrison, H. G. and Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology* **12**, 1889-1898.

Huse, S. M., Ye, Y., Zhou, Y. and Fodor, A. A. (2012). A core human microbiome as viewed through 16S rRNA sequence clusters. *PLoS ONE* **7**, e34242.

Huson, D. H., Mitra, S., Ruscheweyh, H. J., Weber, N. and Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Research* **21**, 1552-1560.

Hutchison III, C. A. (2007). DNA sequencing: Bench to bedside and beyond. *Nucleic Acids Research* **35**, 6227-6237.

Jenkins, M. B., Hartel, P. G., Olexa, T. J. and Stuedemann, J. A. (2003). Putative temporal variability of *Escherichia coli* ribotypes from yearling steers. *Journal of Environmental Quality* **32**, 305-309.

Jenkins, M. W., Tiwari, S., Lorente, M., Gichaba, C. M. and Wuertz, S. (2009). Identifying human and livestock sources of fecal contamination in Kenya with host-specific *Bacteroidales* assays. *Water Research* **43**, 4956-4966.

Jeong, J. Y., Gil, K. I., Lee, K. H. and Ka, J. O. (2008). Molecular identification of fecal pollution sources in water supplies by host-specific fecal DNA markers and terminal restriction fragment length polymorphism profiles of 16S rRNA gene. *Journal of Microbiology* **46**, 599-607.

Jeong, J. Y., Park, H. D., Lee, K. H., Hwang, J. H. and Ka, J. O. (2010). Quantitative analysis of human- and cow-specific 16S rRNA gene markers for assessment of fecal pollution in river waters by real-time PCR. *Journal of Microbiology and Biotechnology* **20**, 245-253.

Jeong, J. Y., Park, H. D., Lee, K. H., Weon, H. Y. and Ka, J. O. (2011). Microbial community analysis and identification of alternative host-specific fecal indicators in fecal and river water samples using pyrosequencing. *Journal of Microbiology* **49**, 585-594.

Jeter, S. N., McDermott, C. M., Bower, P. A., Kinzelman, J. L., Bootsma, M. J., Goetz, G. W. and McLellan, S. L. (2009). *Bacteroidales* diversity in ring-billed gulls (Laurus delawarensis) residing at Lake Michigan beaches. *Applied and Environmental Microbiology* **75**, 1525-1533.

Johnston, C., Ufnar, J. A., Griffith, J. F., Gooch, J. A. and Stewart, J. R. (2010). A real-time qPCR assay for the detection of the *nifH* gene of *Methanobrevibacter smithii*, a potential indicator of sewage pollution. *Journal of Applied Microbiology* **109**, 1946-1956.

Kassa, H., Harrington, B. J. and Bisesi, M. S. (2004). Cryptosporidiosis: A brief literature review and update regarding *Cryptosporidium* in feces of Canada Geese (*Branta canadensis*). *Journal of Environmental Health* **66**, 34-39.

Kembel, S. W., Wu, M., Eisen, J. A. and Green, J. L. (2012). Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Computational Biology* **8**, e1002743.

Khatib, L. A., Tsai, Y. L. and Olson, B. H. (2003). A biomarker for the identification of swine fecal pollution in water, using the STII toxin gene from enterotoxigenic *Escherichia coli*. *Applied Microbiology and Biotechnology* **63**, 231-238.

Kildare, B. J., Leutenegger, C. M., McSwain, B. S., Bambic, D. G., Rajal, V. B. and Wuertz, S. (2007). 16S rRNA-based assays for quantitative detection of universal, human-, cow-, and dog-specific fecal *Bacteroidales*: A Bayesian approach. *Water Research* **41**, 3701-3715.

Kim, M., Morrison, M. and Yu, Z. (2011). Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *Journal of Microbiological Methods* **84**, 81-87.

Kim, O. S., Cho, Y. J., Lee, K., Yoon, S. H., Kim, M., Na, H., Park, S. C., Jeon, Y. S., *et al.* (2012). Introducing EzTaxon-e: A prokaryotic 16s rRNA gene sequence database with phylotypes that represent uncultured species.

International Journal of Systematic and Evolutionary Microbiology **62**, 716-721.

King, E. L., Bachoon, D. S. and Gates, K. W. (2007). Rapid detection of human fecal contamination in estuarine environments by PCR targeting of *Bifidobacterium adolescentis*. *Journal of Microbiological Methods* **68**, 76-81.

Kircher, M. and Kelso, J. (2010). High-throughput DNA sequencing - concepts and limitations. *BioEssays* **32**, 524-536.

Kircher, M., Stenzel, U. and Kelso, J. (2009). Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biology* **10**.

Kirs, M., Harwood, V. J., Fidler, A. E., Gillespie, P. A., Fyfe, W. R., Blackwood, A. D. and Cornelisen, C. D. (2011). Source tracking faecal contamination in an urbanised and a rural waterway in the Nelson-Tasman region, New Zealand. *New Zealand Journal of Marine and Freshwater Research* **45**, 43-58.

Kreader, C. A. (1995). Design and evaluation of Bacteroides DNA probes for the specific detection of human fecal pollution. *Applied and Environmental Microbiology* **61**, 1171-1179.

Krishnan, A. R., Sweeney, M., Vasic, J., Galbraith, D. W. and Vasic, B. (2011). Barcodes for DNA sequencing with guaranteed error correction capability. *Electronics Letters* **47**, 236-237.

Kuczynski, J., Costello, E. K., Nemergut, D. R., Zaneveld, J., Lauber, C. L., Knights, D., Koren, O., Fierer, N., *et al.* (2010). Direct sequencing of the human microbiome readily reveals community differences. *Genome Biology* **11**.

Kuczynski, J., Lauber, C. L., Walters, W. A., Parfrey, L. W., Clemente, J. C., Gevers, D. and Knight, R. (2012). Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics* **13**, 47-58.

Kunin, V., Engelbrektson, A., Ochman, H. and Hugenholtz, P. (2010). Wrinkles in the rare biosphere: Pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology* **12**, 118-123.

Lamendella, R., Santo Domingo, J. W., Ghosh, S., Martinson, J. and Oerther, D. B. (2011). Comparative fecal metagenomics unveils unique functional capacity of the swine gut. *BMC Microbiology* **11**, 103.

Lamendella, R., Santo Domingo, J. W., Kelty, C. and Oerther, D. B. (2008). Bifidobacteria in feces and environmental waters. *Applied and Environmental Microbiology* **74**, 575-584.

Layton, A., McKay, L., Williams, D., Garrett, V., Gentry, R. and Sayler, G. (2006). Development of *Bacteroides* 16S rRNA gene TaqMan-based real-time PCR assays for estimation of total, human, and bovine fecal pollution in water. *Applied and Environmental Microbiology* **72**, 4214-4224.

Lazarevic, V., Whiteson, K., Gaïa, N., Gizard, Y., Hernandez, D., Farinelli, L., Østerås, M., François, P. and Schrenzel, J. (2012). Analysis of the salivary microbiome using culture-independent techniques. *Journal of Clinical Bioinformatics* **2**, e4.

Lazarevic, V., Whiteson, K., Huse, S., Hernandez, D., Farinelli, L., Østerås, M., Schrenzel, J. and François, P. (2009). Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *Journal of Microbiological Methods* **79**, 266-271.

Leamon, J. H., Lee, W. L., Tartaro, K. R., Lanza, J. R., Sarkis, G. J., deWinter, A. D., Berka, J. and Lohman, K. L. (2003). A massively parallel PicoTiterPlate™ based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis* **24**, 3769-3777.

Leclerc, H., Mossel, D. A. A., Edberg, S. C. and Struijk, C. B. (2001). Advances in the bacteriology of the coliform group: Their suitability as markers of microbial water safety. *Annual Review of Microbiology* **55**, 201-234.

Leclerc, H., Schwartzbrod, L. and Dei-Cas, E. (2002). Microbial agents associated with waterborne diseases. *Critical Reviews in Microbiology* **28**, 371-409.

Lee, C. (2011). Genotyping *Escherichia coli* isolates from duck, goose, and gull fecal samples with phylogenetic markers using multiplex polymerase chain reaction for application in microbial source tracking. *Journal of Experimental Microbiology and Immunology* **15**, 130 - 135.

Lee, C. S. and Lee, J. (2010). Evaluation of new *gyrB*-based real-time PCR system for the detection of *B. fragilis* as an indicator of human-specific fecal contamination. *Journal of Microbiological Methods* **82**, 311-318.

Lee, J. E., Lee, S., Sung, J. and Ko, G. (2011). Analysis of human and animal fecal microbiota for microbial source tracking. *ISME Journal* **5**, 362-365.

Leeming, R. and Nichols, P. D. (1996). Concentrations of coprostanol that correspond to existing bacterial indicator guideline limits. *Water Research* **30**, 2997-3006.

Levene, H. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G. and Webb, W. W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**, 682-686.

Ley, R. E., Hamady, M., Lozupone, C., Turnbaugh, P. J., Ramey, R. R., Bircher, J. S., Schlegel, M. L., Tucker, T. A., *et al.* (2008a). Evolution of mammals and their gut microbes. *Science* **320**, 1647-1651.

Ley, R. E., Lozupone, C. A., Hamady, M., Knight, R. and Gordon, J. I. (2008b). Worlds within worlds: Evolution of the vertebrate gut microbiota. *Nature Reviews Microbiology* **6**, 776-788.

Ley, R. E., Peterson, D. A. and Gordon, J. I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124**, 837-848.

Ley, V., Higgins, J. and Fayer, R. (2002). Bovine enteroviruses as indicators of fecal contamination. *Applied and Environmental Microbiology* **68**, 3455-3461.

Li, A., Chu, Y., Wang, X., Ren, L., Yu, J., Liu, X., Yan, J., Zhang, L., *et al.* (2013). A pyrosequencing-based metagenomic study of methane-producing microbial community in solid-state biogas reactor. *Biotechnology for Biofuels* **6**.

Li, W. and Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. and Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology* **2012**, Article 251364.

Liu, Z., Lozupone, C., Hamady, M., Bushman, F. D. and Knight, R. (2007). Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Research* **35**, e120.

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J. and Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology* **30**, 434-439.

Lozupone, C., Hamady, M. and Knight, R. (2006). UniFrac - An online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7**, e371.

Lozupone, C. and Knight, R. (2005). UniFrac: A new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* **71**, 8228-8235.

Lozupone, C. A., Hamady, M., Kelley, S. T. and Knight, R. (2007). Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology* **73**, 1576-1585.

Lozupone, C. A. and Knight, R. (2007). Global patterns in bacterial diversity. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 11436-11440.

Lozupone, C. A. and Knight, R. (2008). Species divergence and the measurement of microbial diversity. *FEMS Microbiology Reviews* **32**, 557-578.

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. and Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220-230.

Lu, J. and Domingo, J. S. (2008). Turkey fecal microbial community structure and functional gene diversity revealed by 16S rRNA gene and metagenomic sequences. *Journal of Microbiology* **46**, 469-477.

Lu, J., Ryu, H., Hill, S., Schoen, M., Ashbolt, N., Edge, T. A. and Domingo, J. S. (2011). Distribution and potential significance of a gull fecal marker in urban coastal and riverine areas of southern Ontario, Canada. *Water Research* **45**, 3960-3968.

Lu, J., Santo Domingo, J. and Shanks, O. C. (2007). Identification of chicken-specific fecal microbial sequences using a metagenomic approach. *Water Research* **41**, 3561-3574.

Lu, J., Santo Domingo, J. W., Lamendella, R., Edge, T. and Hill, S. (2008). Phylogenetic diversity and molecular detection of bacteria in gull feces. *Applied and Environmental Microbiology* **74**, 3969-3976.

Lu, L., Hume, M. E., Sternes, K. L. and Pillai, S. D. (2004). Genetic diversity of *Escherichia coli* isolates in irrigation water and associated sediments: Implications for source tracking. *Water Research* **38**, 3899-3908.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, A., Buchner, A., Lai, T., *et al.* (2004). ARB: A software environment for sequence data. *Nucleic Acids Research* **32**, 1363-1371.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics* **9**, 387-402.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380.

Martellini, A., Payment, P. and Villemur, R. (2005). Use of eukaryotic mitochondrial DNA to differentiate human, bovine, porcine and ovine sources in fecally contaminated surface water. *Water Research* **39**, 541-548.

Marti, R., Dabert, P., Ziebal, C. and Pourcher, A. M. (2010). Evaluation of *Lactobacillus sobrius*/*L. amylovorus* as a new microbial marker of pig manure. *Applied and Environmental Microbiology* **76**, 1456-1461.

Matsuki, T., Watanabe, K., Fujimoto, J., Kado, Y., Takada, T., Matsumoto, K. and Tanaka, R. (2004). Quantitative PCR with 16S rRNA-gene-targeted species-specific primers for analysis of human intestinal Bifidobacteria. *Applied and Environmental Microbiology* **70**, 167-173.

McBride, G. B., Salmond, C. E., Bandaranayake, D. R., Turner, S. J., Lewis, G. D. and Till, D. G. (1998). Health effects of marine bathing in New Zealand. *International Journal of Environmental Health Research* **8**, 173-189.

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., Desantis, T. Z., Probst, A., Andersen, G. L., Knight, R. and Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME Journal* **6**, 610-618.

McLain, J. E. T., Ryu, H., Kabiri-Badr, L., Rock, C. M. and Abbaszadegan, M. (2009). Lack of specificity for PCR assays targeting human *Bacteroides* 16S rRNA gene: Cross-amplification with fish feces. *FEMS Microbiology Letters* **299**, 38-43.

McLellan, S. L. (2004). Genetic diversity of *Escherichia coli* isolated from urban rivers and beach water. *Applied and Environmental Microbiology* **70**, 4658-4665.

McLellan, S. L., Daniels, A. D. and Salmore, A. K. (2001). Clonal populations of thermotolerant *Enterobacteriaceae* in recreational water and their potential interference with fecal *Escherichia coli c*ounts. *Applied and Environmental Microbiology* **67**, 4934-4938.

McLellan, S. L., Daniels, A. D. and Salmore, A. K. (2003). Genetic characterization of *Escherichia coli* populations from host sources of fecal pollution by using DNA fingerprinting. *Applied and Environmental Microbiology* **69**, 2587-2594.

McLellan, S. L., Huse, S. M., Mueller-Spitz, S. R., Andreishcheva, E. N. and Sogin, M. L. (2010). Diversity and population structure of sewage-derived microorganisms in wastewater treatment plant influent. *Environmental Microbiology* **12**, 378-392.

McQuaig, S. M. and Noble, R. T. (2011). Virses as tracers of fecal contamination. In *Microbial source tracking: Methods, applications and case studies*, pp. 113-136. Edited by C. Hagedorn, V. J. Harwood and A. R. Blanch. New York, USA: Springer Science+Business Media.

McQuaig, S. M., Scott, T. M., Harwood, V. J., Farrah, S. R. and Lukasik, J. O. (2006). Detection of human-derived fecal pollution in environmental waters by use of a PCR-based human polyomavirus assay. *Applied and Environmental Microbiology* **72**, 7567-7574.

Meays, C. L., Broersma, K., Nordin, R. and Mazumder, A. (2004). Source tracking fecal bacteria in water: A critical review of current methods. *Journal of Environmental Management* **73**, 71-79.

Meays, C. L., Broersma, K., Nordin, R., Mazumder, A. and Samadpour, M. (2006). Spatial and annual variability in concentrations and sources of *Escherichia coli* in multiple watersheds. *Environmental Science and Technology* **40**, 5289-5296.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics* **11**, 31-46.

Meyer, M., Stenzel, U. and Hofreiter, M. (2008). Parallel tagged sequencing on the 454 platform. *Nature Protocols* **3**, 267-278.

Meyer, M., Stenzel, U., Myles, S., Prüfer, K. and Hofreiter, M. (2007). Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Research* **35**.

Midura, T. F. and Bryant, R. G. (2001). Sample plans, sample collection, shipment and preparation for analysis. In *Compendium of methods for the microbiological examination of foods*, pp. 13-23. Edited by F. P. I. Downes, K. Washington, DC, USA: American Public Health Association.

Mieszkin, S., Furet, J. P., Corthier, G. and Gourmelon, M. (2009). Estimation of pig fecal contamination in a river catchment by real-time PCR using two Pig-Specific *Bacteroidales* 16S rRNA genetic markers. *Applied and Environmental Microbiology* **75**, 3045-3054.

Mieszkin, S., Yala, J. F., Joubrel, R. and Gourmelon, M. (2010). Phylogenetic analysis of *Bacteroidales* 16S rRNA gene sequences from human and animal effluents and assessment of ruminant faecal pollution by real-time PCR. *Journal of Applied Microbiology* **108**, 974-984.

Mizrahi-Man, O., Davenport, E. R. and Gilad, Y. (2013). Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: Evaluation of effective study designs. *PLoS ONE* **8**, e53608.

Moorthie, S., Mattocks, C. J. and Wright, C. F. (2011). Review of massively parallel DNA sequencing technologies. *HUGO Journal* **5**, 1-12.

Moriarty, E. M., Karki, N., MacKenzie, M., Sinton, L. W., Wood, D. R. and Gilpin, B. J. (2011). Faecal indicators and pathogens in selected New Zealand waterfowl. *New Zealand Journal of Marine and Freshwater Research* **45**, 679-688.

Mott, J. and Smith, A. (2011). Library-dependent source tracking methods. In *Microbial source tracking: Methods, applications and case studies*, pp. 31-60. Edited by C. Hagedorn, V. J. Harwood and A. R. Blanch. New York, USA: Springer Science+Business Media.

Myoda, S. P., Carson, C. A., Fuhrmann, J. J., Hahm, B. K., Hartel, P. G., Yampara-Lquise, H., Johnson, L., Kuntz, R. L., *et al.* (2003). Comparison of genotypic-based microbial source tracking methods requiring a host origin database. *Journal of Water and Health* **1**, 167-180.

Nawrocki, E. P., Kolbe, D. L. and Eddy, S. R. (2009). Infernal 1.0: Inference of RNA alignments. *Bioinformatics* **25**, 1335-1337.

Noble, R. T., Allen, S. M., Blackwood, A. D., Chu, W., Jiang, S. C., Lovelace, G. L., Sobsey, M. D., Stewart, J. R. and Wait, D. A. (2003). Use of viral pathogens and indicators to differentiate between human and non-human fecal contamination in a microbial source tracking comparison study. *Journal of Water and Health* **1**, 195-207.

Okabe, S., Okayama, N., Savichtcheva, O. and Ito, T. (2007). Quantification of host-specific *Bacteroides-Prevotella* 16S rRNA genetic markers for assessment of fecal pollution in freshwater. *Applied Microbiology and Biotechnology* **74**, 890-901.

Okabe, S. and Shimazu, Y. (2007). Persistence of host-specific *Bacteroides-Prevotella* 16S rRNA genetic markers in environmental waters: Effects of temperature and salinity. *Applied Microbiology and Biotechnology* **76**, 935-944.

Opel, K. L., Chung, D. and McCord, B. R. (2010). A study of PCR inhibition mechanisms using real time PCR. *Journal of Forensic Sciences* **55**, 25-33.

Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., McInerney, P., Thompson, J. F., Bowers, J., *et al.* (2009). Direct RNA sequencing. *Nature* **461**, 814-818.

Parameswaran, P., Jalili, R., Tao, L., Shokralla, S., Gharizadeh, B., Ronaghi, M. and Fire, A. Z. (2007). A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Research* **35**, e130.

Pareek, C. S., Smoczynski, R. and Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics* **52**, 413-435.

Parsley, L. C., Consuegra, E. J., Thomas, S. J., Bhavsar, J., Land, A. M., Bhuiyan, N. N., Mazher, M. A., Waters, R. J., *et al.* (2010). Census of the viral metagenome within an activated sludge microbial assemblage. *Applied and Environmental Microbiology* **76**, 2673-2677.

Peeler, K. A., Opsahl, S. P. and Chanton, J. P. (2006). Tracking anthropogenic inputs using caffeine, indicator bacteria, and nutrients in rural freshwater and urban marine systems. *Environmental Science and Technology* **40**, 7616-7622.

Plummer, J. D. and Long, S. C. (2009). Identifying sources of surface water pollution: A toolbox approach. *Journal - American Water Works Association* **101**, 75-88, 16.

Price, M. N., Dehal, P. S. and Arkin, A. P. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490.

Prüss, A. (1998). Review of epidemiological studies on health effects from exposure to recreational water. *International Journal of Epidemiology* **27**, 1-9.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., *et al.* (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65.

Quail, M. A., Otto, T. D., Gu, Y., Harris, S. R., Skelly, T. F., McQuillan, J. A., Swerdlow, H. P. and Oyola, S. O. (2012a). Optimal enzymes for amplifying sequencing libraries. *Nature Methods* **9**, 10-11.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P. and Gu, Y. (2012b). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, e341.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research* **41**, D590-D596.

Quinlan, A. R., Stewart, D. A., Strömberg, M. P. and Marth, G. T. (2008). Pyrobayes: An improved base caller for SNP discovery in pyrosequences. *Nature Methods* **5**, 179-181.

Ram, J. L., Thompson, B., Turner, C., Nechvatal, J. M., Sheehan, H. and Bobrin, J. (2007). Identification of pets and raccoons as sources of bacterial contamination of urban storm sewers using a sequence-based bacterial source tracking method. *Water Research* **41**, 3605-3614.

Ratan, A., Miller, W., Guillory, J., Stinson, J., Seshagiri, S. and Schuster, S. C. (2013). Comparison of sequencing platforms for single nucleotide variant calls in a human sample. *PLoS ONE* **8**, e55089.

Reischer, G. H., Kasper, D. C., Steinborn, R., Farnleitner, A. H. and Mach, R. L. (2007). A quantitative real-time PCR assay for the highly sensitive and specific detection of human faecal influence in spring water from a large alpine catchment area. *Letters in Applied Microbiology* **44**, 351-356.

Reischer, G. H., Kasper, D. C., Steinborn, R., Mach, R. L. and Farnleitner, A. H. (2006). Quantitative PCR method for sensitive detection of ruminant fecal pollution in freshwater and evaluation of this method in alpine karstic regions. *Applied and Environmental Microbiology* **72**, 5610-5614.

Ritter, K. J., Carruthers, E., Carson, C. A., Ellender, R. D., Harwood, V. J., Kingsley, K., Nakatsu, C., Sadowsky, M., *et al.* (2003). Assessment of statistical methods used in library-based approaches to microbial source tracking. *Journal of Water and Health* **1**, 209-223.

Roberts, P. H. and Thomas, K. V. (2006). The occurrence of selected pharmaceuticals in wastewater effluent and surface waters of the lower Tyne catchment. *Science of the Total Environment* **356**, 143-153.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. and Nyrén, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry* **242**, 84-89.

Ronaghi, M., Uhlén, M. and Nyrén, P. (1998). A sequencing method based on real-time pyrophosphate. *Science* **281**, 363-365.

Roossinck, M. J., Saha, P., Wiley, G. B., Quan, J., White, J. D., Lai, H., Chavarría, F., Shen, G. and Roe, B. A. (2010). Ecogenomics: Using massively parallel pyrosequencing to understand virus ecology. *Molecular Ecology* **19**, 81-88.

Roslev, P. and Bukh, A. S. (2011). State of the art molecular markers for fecal pollution source tracking in water. *Applied Microbiology and Biotechnology* **89**, 1341-1355.

Rossello-Mora, R. A., Wagner, M., Amann, R. and Schleifer, K. H. (1995). The abundance of *Zoogloea ramigera* in sewage treatment plants. *Applied and Environmental Microbiology* **61**, 702-707.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., *et al.* (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348-352.

Rothberg, J. M. and Leamon, J. H. (2008). The development and impact of 454 sequencing. *Nature Biotechnology* **26**, 1117-1124.

Sanapareddy, N., Hamp, T. J., Gonzalez, L. C., Hilger, H. A., Fodor, A. A. and Clinton, S. M. (2009). Molecular diversity of a North Carolina wastewater treatment plant as revealed by pyrosequencing. *Applied and Environmental Microbiology* **75**, 1688-1696.

Sanger, F., Air, G. M. and Barrell, B. G. (1977a). Nucleotide sequence of bacteriophage ΦX174 DNA. *Nature* **265**, 687-695.

Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* **94**, 441-448.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977b). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5463-5467.

Santo Domingo, J. W., Bambic, D. G., Edge, T. A. and Wuertz, S. (2007). Quo vadis source tracking? Towards a strategic framework for environmental monitoring of fecal pollution. *Water Research* **41**, 3539-3552.

Saunders, A. M., Kristiansen, A., Lund, M. B., Revsbech, N. P. and Schramm, A. (2009). Detection and persistence of fecal *Bacteroidales* as water quality indicators in unchlorinated drinking water. *Systematic and Applied Microbiology* **32**, 362-370.

Savichtcheva, O., Okayama, N. and Okabe, S. (2007). Relationships between *Bacteroides* 16S rRNA genetic markers and presence of bacterial enteric pathogens and conventional fecal indicators. *Water Research* **41**, 3615-3628.

Savill, M. G., Murray, S. R., Scholes, P., Maas, E. W., McCormick, R. E., Moore, E. B. and Gilpin, B. J. (2001). Application of polymerase chain reaction (PCR) and TaqMan™ PCR techniques to the detection and identification of *Rhodococcus coprophilus* in faecal samples. *Journal of Microbiological Methods* **47**, 355-368.

Schadt, E. E., Turner, S. and Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics* **19**, R227-R240.

Schloss, P. D., Gevers, D. and Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16s rRNA-based studies. *PLoS ONE* **6**, e27310.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., *et al.* (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75**, 7537-7541.

Scott, T. M., Jenkins, T. M., Lukasik, J. and Rose, J. B. (2005). Potential use of a host associated molecular marker in Enterococcus faecium as an index of human fecal pollution. *Environmental Science and Technology* **39**, 283-287.

Scott, T. M., Rose, J. B., Jenkins, T. M., Farrah, S. R. and Lukasik, J. (2002). Microbial source tracking: Current methodology and future directions. *Applied and Environmental Microbiology* **68**, 5796-5803.

Seurinck, S., Defoirdt, T., Verstraete, W. and Siciliano, S. D. (2005). Detection and quantification of the human-specific HF183 *Bacteroides* 16S rRNA genetic marker with real-time PCR for assessment of human faecal pollution in freshwater. *Environmental Microbiology* **7**, 249-259.

Shanks, O. C., Atikovic, E., Blackwood, A. D., Lu, J., Noble, R. T., Domingo, J. S., Seifring, S., Sivaganesan, M. and Haugland, R. A. (2008). Quantitative PCR for detection and enumeration of genetic markers of bovine fecal pollution. *Applied and Environmental Microbiology* **74**, 745-752.

Shanks, O. C., Domingo, J. W. S., Lu, J., Kelty, C. A. and Graham, J. E. (2007). Identification of bacterial DNA markers for the detection of human fecal pollution in water. *Applied and Environmental Microbiology* **73**, 2416.

Shanks, O. C., Kelty, C. A., Sivaganesan, M., Varma, M. and Haugland, R. A. (2009). Quantitative PCR for genetic markers of human fecal pollution. *Applied and Environmental Microbiology* **75**, 5507-5513.

Shanks, O. C., Santo Domingo, J. W., Lamendella, R., Kelty, C. A. and Graham, J. E. (2006). Competitive metagenomic DNA hybridization identifies host-specific microbial genetic markers in cow fecal samples. *Applied and Environmental Microbiology* **72**, 4054.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., *et al.* (2005). Molecular biology: Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-1732.

Shibata, T., Solo-Gabriele, H. M., Sinigalliano, C. D., Gidley, M. L., Plano, L. R. W., Fleisher, J. M., Wang, J. D., Elmir, S. M., *et al.* (2010). Evaluation of conventional and alternative monitoring methods for a recreational marine beach with nonpoint source of fecal contamination. *Environmental Science and Technology* **44**, 8175-8181.

Shokralla, S., Spall, J. L., Gibson, J. F. and Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology* **21**, 1794-1805.

Shuval, H. (2003). Estimating the global burden of thalassogenic diseases: human infectious diseases caused by wastewater pollution of the marine environment. *Journal of Water and Health* **1**, 53-64.

Siefring, S., Varma, M., Atikovic, E., Wymer, L. and Haugland, R. A. (2008). Improved real-time PCR assays for the detection of fecal indicator bacteria in

surface waters with different instrument and reagent systems. *Journal of Water and Health* **6**, 225-237.

Silkie, S. S. and Nelson, K. L. (2009). Concentrations of host-specific and generic fecal markers measured by quantitative PCR in raw sewage and fresh animal feces. *Water Research* **43**, 4860-4871.

Simpson, J. M., Santo Domingo, J. W. and Reasoner, D. J. (2002). Microbial source tracking: State of the science. *Environmental Science and Technology* **36**, 5279-5288.

Sinton, L. W., Finlay, R. K. and Hannah, D. J. (1998). Distinguishing human from animal faecal contamination in water: A review. *New Zealand Journal of Marine and Freshwater Research* **32**, 323-348.

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M. and Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America* **103**, 12115-12120.

Soule, M., Kuhn, E., Loge, F., Gay, J. and Call, D. R. (2006). Using DNA microarrays to identify library-independent markers for bacterial source tracking. *Applied and Environmental Microbiology* **72**, 1843-1851.

Stewart, J. R., Ellender, R. D., Gooch, J. A., Jiang, S., Myoda, S. P. and Weisberg, S. B. (2003). Recommendations for microbial source tracking: Lessons from a methods comparison study. *Journal of Water and Health* **1**, 225-231.

Stewart, J. R., Santo Domingo, J. W. and Wade, T. J. (2007). Fecal pollution, public health and microbial source tracking. In *Microbial source tracking*, pp. 1-32. Edited by J. W. Santo Domingo and M. J. Sadowsky. Washington, DC, USA: ASM Press.

Stoddart, D., Heron, A. J., Mikhailova, E., Maglia, G. and Bayley, H. (2009). Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 7702-7707.

Stoeckel, D. M. and Harwood, V. J. (2007). Performance, design, and analysis in microbial source tracking studies. *Applied and Environmental Microbiology* **73**, 2405-2415.

Stoeckel, D. M., Mathes, M. V., Hyer, K. E., Hagedorn, C., Kator, H., Lukasik, J., O'Brien, T. L., Fenger, T. W., *et al.* (2004). Comparison of seven protocols to identify fecal contamination sources using *Escherichia coli*. *Environmental Science and Technology* **38**, 6109-6117.

Stricker, A. R., Wilhartitz, I., Farnleitner, A. H. and Mach, R. L. (2008). Development of a Scorpion probe-based real-time PCR for the sensitive quantification of *Bacteroides* sp. ribosomal DNA from human and cattle origin and evaluation in spring water matrices. *Microbiological Research* **163**, 140-147.

Su, C., Lei, L., Duan, Y., Zhang, K. Q. and Yang, J. (2012). Culture-independent methods for studying environmental microorganisms: Methods, application, and perspective. *Applied Microbiology and Biotechnology* **93**, 993-1003.

Suchodolski, J. S., Camacho, J. and Steiner, J. M. (2008). Analysis of bacterial diversity in the canine duodenum, jejunum, ileum, and colon by comparative 16S rRNA gene analysis. *FEMS Microbiology Ecology* **66**, 567-578.

Suzuki, M., Rappé, M. S. and Giovannoni, S. J. (1998). Kinetic bias in estimates of coastal picoplankton community structure obtained by measurements of small-

subunit rRNA gene PCR amplicon length heterogeneity. *Applied and Environmental Microbiology* **64**, 4522-4529.

Suzuki, S., Ono, N., Furusawa, C., Ying, B. W. and Yomo, T. (2011). Comparison of sequence reads obtained from three next-generation sequencing platforms. *PLoS ONE* **6**, e19534.

Tajima, K., Aminov, R. I., Nagamine, T., Matsui, H., Nakamura, M. and Benno, Y. (2001). Diet-dependent shifts in the bacterial population of the rumen revealed with real-time PCR. *Applied and Environmental Microbiology* **67**, 2766-2774.

Tallon, P., Magajna, B., Lofranco, C. and Kam, T. L. (2005). Microbial indicators of faecal contamination in water: A current perspective. *Water, Air, and Soil Pollution* **166**, 139-166.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**, 2731-2739.

Teaf, C. M. and Garber, M. M. (2011). Use of microbial source tracking in the legal arena: Benefits and challenges. In *Microbial source tracking: Methods, applications and case studies*, pp. 301-312. Edited by C. Hagedorn, A. R. Blanch and V. J. Harwood. New York, USA: Springer Science+Business Media.

Thomas, T., Gilbert, J. and Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation* **2**, 3-14.

Tripathi, B. M., Kim, M., Singh, D., Lee-Cruz, L., Lai-Hoe, A., Ainuddin, A. N., Go, R., Rahim, R. A., *et al.* (2012). Tropical soil bacterial communities in Malaysia: pH dominates in the equatorial tropics too. *Microbial Ecology* **64**, 474-484.

Tsuchiya, C., Sakata, T. and Sugita, H. (2008). Novel ecological niche of *Cetobacterium somerae*, an anaerobic bacterium in the intestinal tracts of freshwater fish. *Letters in Applied Microbiology* **46**, 43-48.

Turcatti, G., Romieu, A., Fedurco, M. and Tairi, A. P. (2008). A new class of cleavable fluorescent nucleotides: Synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Research* **36**, e25.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., *et al.* (2009). A core gut microbiome in obese and lean twins. *Nature* **457**, 480-484.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R. and Gordon, J. I. (2007). The Human Microbiome Project. *Nature* **449**, 804-810.

Unno, T., Jang, J., Han, D., Kim, J. H., Sadowsky, M. J., Kim, O. S., Chun, J. and Hur, H. G. (2010). Use of barcoded pyrosequencing and shared OTUs to determine sources of fecal bacteria in watersheds. *Environmental Science and Technology* **44**, 7777-7782.

Valverde, J. R. and Mellado, R. P. (2013). Analysis of metagenomic data containing high biodiversity levels. *PLoS ONE* **8**, e58118.

Venter, C. J., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., *et al.* (2001). The sequence of the human genome. *Science* **291**, 1304-1351.

Verhulst, N. O., Qiu, Y. T., Beijleveld, H., Maliepaard, C., Knights, D., Schulz, S., Berg-Lyons, D., Lauber, C. L., *et al.* (2011). Composition of human skin microbiota affects attractiveness to malaria mosquitoes. *PLoS ONE* **6**, e28991.

Větrovský, T. and Baldrian, P. (2013). The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS ONE* **8**, e57923.

Vilo, C. and Dong, Q. (2012). Evaluation of the RDP Classifier accuracy using 16S rRNA gene variable regions. *Metagenomics* **1**, 1-5.

Voelkerding, K. V., Dames, S. A. and Durtschi, J. D. (2009). Next-generation sequencing: From basic research to diagnostics. *Clinical Chemistry* **55**, 641-658.

Vogel, J. R., Stoeckel, D. M., Lamendella, R., Zelt, R. B., Santo Domingo, J. W., Walker, S. R. and Oerther, D. B. (2007). Identifying fecal sources in a selected catchment reach using multiple source-tracking tools. *Journal of Environmental Quality* **36**, 718-729.

Walters, S. P. and Field, K. G. (2009). Survival and persistence of human and ruminant-specific faecal *Bacteroidales* in freshwater microcosms. *Environmental Microbiology* **11**, 1410-1421.

Wang, D., Silkie, S. S., Nelson, K. L. and Wuertz, S. (2010). Estimating true human and animal host source contribution in quantitative microbial source tracking using the Monte Carlo method. *Water Research* **44**, 4760-4775.

Wang, Q., Garrity, G. M., Tiedje, J. M. and Cole, J. R. (2007). Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* **73**, 5261-5267.

Wang, R. F., Beggs, M. L., Erickson, B. D. and Cerniglia, C. E. (2004). DNA microarray analysis of predominant human intestinal bacteria in fecal samples. *Molecular and Cellular Probes* **18**, 223-234.

Wicker, T., Schlagenhauf, E., Graner, A., Close, T. J., Keller, B. and Stein, N. (2006). 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* **7**, e275.

Wiggins, B. A., Cash, P. W., Creamer, W. S., Dart, S. E., Garcia, P. P., Gerecke, T. M., Han, J., Henry, B. L., *et al.* (2003). Use of antibiotic resistance analysis for representativeness testing of multiwatershed libraries. *Applied and Environmental Microbiology* **69**, 3399-3405.

Wilhelm, S. W., Farnsley, S. E., LeCleir, G. R., Layton, A. C., Satchwell, M. F., DeBruyn, J. M., Boyer, G. L., Zhu, G. and Paerl, H. W. (2011). The relationships between nutrients, cyanobacterial toxins and the microbial community in Taihu (Lake Tai), China. *Harmful Algae* **10**, 207-215.

Wooley, J. C., Godzik, A. and Friedberg, I. (2010). A primer on metagenomics. *PLoS Computational Biology* **6**, e10000667.

Wu, G. D., Lewis, J. D., Hoffmann, C., Chen, Y. Y., Knight, R., Bittinger, K., Hwang, J., Chen, J., *et al.* (2010). Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. *BMC Microbiology* **10**, e206.

Wuertz, S., Wang, D., Reischer, G. H. and Farnleitner, A. H. (2011). Library-independent bacterial source tracking methods. In *Microbial source tracking: Methods, applications and case studies*, pp. 61-112. Edited by C. Hagedorn, V. J. Harwood and A. R. Blanch. New York, USA: Springer Science+Business Media.

Yamahara, K. M., Layton, B. A., Santoro, A. E. and Boehm, A. B. (2007). Beach sands along the California coast are diffuse sources of fecal bacteria to coastal waters. *Environmental Science and Technology* **41**, 4515-4521.

Yampara-Iquise, H., Zheng, G., Jones, J. E. and Carson, C. A. (2008). Use of a *Bacteroides thetaiotaomicron*-specific α-1-6, mannanase quantitative PCR to

detect human faecal pollution in water. *Journal of Applied Microbiology* **105**, 1686-1693.

Yan, T. and Sadowsky, M. J. (2007). Determining sources of fecal bacteria in waterways. *Environmental Monitoring and Assessment* **129**, 97-106.

Ye, L., Zhang, T., Wang, T. and Fang, Z. (2012). Microbial structures, functions, and metabolic pathways in wastewater treatment bioreactors revealed using high-throughput sequencing. *Environmental Science and Technology* **46**, 13244-13252.

Zarraonaindia, I., Smith, D. P. and Gilbert, J. A. (2013). Beyond the genome: Community-level analysis of the microbial world. *Biology and Philosophy* **28**, 261-282.

Zheng, G., Yampara-Iquise, H., Jones, J. E. and Carson, C. A. (2009). Development of *Faecalibacterium* 16S rRNA gene marker for identification of human faeces. *Journal of Applied Microbiology* **106**, 634-641.

Zhou, J., Wu, L., Deng, Y., Zhi, X., Jiang, Y. H., Tu, Q., Xie, J., Van Nostrand, J. D., *et al.* (2011). Reproducibility and quantitation of amplicon sequencing-based detection. *ISME Journal* **5**, 1303-1313.

Zhou, X., Ren, L., Li, Y., Zhang, M., Yu, Y. and Yu, J. (2010a). The next-generation sequencing technology: A technology review and future perspective. *Science China Life Sciences* **53**, 44-57.

Zhou, X., Ren, L., Meng, Q., Li, Y., Yu, Y. and Yu, J. (2010b). The next-generation sequencing technology and application. *Protein and Cell* **1**, 520-536.

Appendix I

QIIME scripts used during analysis

All required files were placed within a working folder, to ensure pathfiles were the same. This includes alignment and lanemask files initially downloaded during the setup of QIIME.

General parameter codes

-i, name and file path for input file required for analysis

-o, name and file path for output directory

-m, name and file path for metadata mapping file

-a –O x, allows workflow scripts to be run in parallel, for use in a multi-core or cluster environment. The value of x defines how many jobs to run simultaneously, which will depend on the computer being used. As the analysis in this study was being run on an 8 processor machine, with other users also potentially using it, a value of 4 was generally used.

-e, allows the user to define the sampling depth for diversity analysis, ensuring samples are only sampled to a depth that is suitable for all samples. QIIME suggests this value to be the smallest number of OTUs assigned to a sample for calculating alpha and beta diversity, and 75% of the smallest number of OTUs for calculating jackknifed replicates.

GS454-01A sequencing data (Chapter 2)

Filtering and demultiplexing of data

```
$check_id_map.py –m GS45401_Metadata_forward.txt –o
Forward_map_output/
```

```
$check_id_map.py –m GS45401_Metadata_reverse.txt –o
Reverse_map_output/
```

Ensures user-created mapping files are formatted correctly.

```
$split_libraries.py –m GS45401_Metadata_forward.txt –f
Amplicon01_reads.
fasta –q Amplicon01_reads.qual –o 01_Forward_library_output/ -b 4 –z
truncate_only

$split_libraries.py –m GS45401_Metadata_reverse.txt –f
Amplicon01_reads.
fasta –q Amplicon01_reads.qual –o 01_Reverse_library_output/ -b 4 –z
truncate_only –n 2551
```

Separates sequences based on information provided in the mapping file, renaming each sequence based on its appropriate sample ID. Required input files are a sequence FASTA file and related quality file, as well as a metadata mapping file. A large number of filtering options are also included, which can be controlled by the user. The defaults include minimum sequence length of 200, maximum sequence length of 1000, a minimum quality score of 25, a maximum of 6 ambiguous bases, a maximum of 6 homopolymers, and no primer mismatches. The forward primer and barcode sequences are also removed. The barcode type can be defined using the –b parameter, with the type or length of barcode included. The –z parameter allows for the removal of the reverse primer sequence, and any following nucleotides. The –n parameter allows the user to define the starting number for sequence ID. This ensures that if multiple files are to be combined, no sequence will have the same sequence ID number.

```
$adjust_seq_orientation.py –i 01_Reverse_library_output/seqs.fna
```

Writes the reverse complements all the sequences in the file.

```
$cat 01_Forward_library_output/seqs.fna 01_Reverse_library_output/
seqs_rc.fna >01_ Combined seqs.fna
```

Concatenates multiple sequence files together, to form a single file for further analysis.

Picking Operational Taxonomic Units (OTUs)

```
$pick_otus_through_otu_table.py –i 01_Combined_seqs.fna –o 01_otus/
```

A workflow script that includes a number of steps to pick OTUs, align and filter representative sequences for each OTU, assign taxonomy to each OTU, create a

phylogenetic tree and assemble an OTU table required for downstream analysis. Scripts included in the workflow are:

```
$pick_otus.py
$pick_rep_set.py
$align_seqs.py
$filter_alignment.py
$assign_taxonomy.py
$make_phylogeny.py
$make_otu_table.py
```

```
$per_library_stats.py –i 01_otus/otu_table.biom
```

Provides details on the number of sequence reads assigned to each sample within the OTU table.

```
$summarize_taxa_through_plots.py –i 01_otus/otu_table.biom –o
01_taxa_summary –m GS45401_Metadata_forward.txt
```

```
$summarize_taxa_through_plots.py –i 01_otus/otu_table.biom –o
01_species_taxa_summary –m GS45401_Metadata_forward.txt –c Species
```

Groups OTUs by samples or categories at the different taxonomic levels. Different categories can be used based on the information provided in the metadata file, using the –c parameter. Outputs include tables for each taxonomic level and html files for area and bar charts.

Alpha diversity

```
$echo “alpha_diversity: metrics Shannon,PD_whole_tree,chao1,observed_
species” > alpha_params.txt
```

Creates a custom parameter file allowing the user to define which diversity metrics to be included in analysis.

```
$alpha_rarefaction.py –i 01_otus/otu_table.biom –m
GS45401_Metadata_forward.txt –o 01_alpha_rare/ -p alpha_params.txt –t
01_otus/rep_set.tre –e 116
```

A workflow script that determines the alpha diversity in the samples by generating rarefied OTU tables, computing measures of alpha diversity for each rarefied OTU table, collates alpha diversity results and generated alpha rarefaction plots.

A phylogenetic tree is required if phylogenetic metrics, such as PD_whole_tree are included in the analyses. The inclusion of the –p parameter allows the user to define which alpha diversity metrics are calculated. The –e value used here was the smallest number of OTUs assigned to a single sample. Scripts included in the workflow are:

```
$multiple_rarefactions.py
$alpha_diversity.py
$collate_alpha.py
$make_rarefaction_plots.py
```

Beta diversity

```
$beta_diversity_through_plots.py –i 01_otus/otu_table.biom –m GS45401_
Metadata_forward.txt –o 01_beta_div/ -t 01_otus/rep_set.tre –e 116
```

A workflow script that determines beta diversity in the samples by generating rarefied samples, computing beta diversity, generating Principal Coordinates and generating 2D and 3D plots. The default metrics are weighted and unweighted UniFrac, which can be changed by the addition of a custom parameter file if required. As weighted UniFrac is a phylogenetic measure, a phylogenetic tree is required. Scripts included in the workflow are:

```
$single_rarefaction.py
$make_prefs_file.py
$beta_diversity.py
$principal_coordinates.py
$make_3d_plots.py
$make_2d_plots.py
```

```
$jackknifed_beta_diversity.py –i 01_otus/otu_table.biom –t rep_set.tre
–m GS45401_Metadata_forward.txt –o jack_div/ -e 87
```

A workflow script that uses jackknifed replicates to estimate uncertainty in PCoA plots and hierarchical clustering of microbial communities, through computing UPGMA clustering of a set number of sequences from each sample, generating jackknife replicates, comparing jackknifed trees and creating jackknifed supporting trees, and comparing PCoA plots.

Combined sequencing data (Chapter 3)

Filtering and demultiplexing of data

```
$split_libraries.py –m GS45401_Metadata_forward.txt –f
Amplicon01B_reads.
fasta –q Amplicon01B_reads.qual –o 01B_Forward_library_output/ -b 4 –z
truncate_only

$split_libraries.py –m GS45401_Metadata_reverse.txt –f
Amplicon01B_reads.
fasta –q Amplicon01B_reads.qual –o 01B_Reverse_library_output/ -b 4 –z
truncate_only –n 71000

$adjust_seq_orientation.py –i 01B_Reverse_library_output/seqs.fna

$cat 01B_Forward_library_output/seqs.fna 01B_Reverse_library_output/
seqs_rc.fna > 01B_Combined_seqs.fna

$extract_seqs_by_sample_id.py –i 01B_Combined_seqs.fna –o
01B_seqs_by_sample.fasta –s NGS003.B4.3,NGS005.B4.5,NGS019.B4.19,
NGS020.B4.20 –n
```

Allows the removal of sequences associated with certain sample IDs. The –n option keeps all sequences that are not associated with the supplied sample IDs.

```
$split_libraries.py –m GS45402_Metadata_forward.txt –f
Amplicon02_reads.
fasta –q Amplicon02_reads.qual –o 02_Forward_library_output/ -b
hamming_8 –z truncate_only –n 155000

$split_libraries.py –m GS45402_Metadata_reverse.txt –f
Amplicon02_reads.
fasta –q Amplicon02_reads.qual –o 02_Reverse_library_output/ -b
hamming_8 –z truncate_only –n 230000

$adjust_seq_orientation.py –i 02_Reverse_library_output/seqs.fna

$cat 02_Forward_library_output/seqs.fna 02_Reverse_library_output/
seqs_rc.fna > 02_Combined_seqs.fna

$split_libraries.py –m GS45403_Metadata_forward.txt –f
Amplicon03_reads.
fasta –q Amplicon03_reads.qual –o 03_Forward_library_output/ -b
hamming_8 –z truncate_only –n 290000
```

```
$split_libraries.py –m GS45403_Metadata_forward.txt –f
Amplicon03_reads.
fasta –q Amplicon03_reads.qual –o 03_Reverse_library_output/ -b
hamming_8 –z truncate_only –n 350000

$adjust_seq_orientation.py –i 03_Reverse_library_output/seqs.fna

$cat 03_Forward_library_output/seqs.fna 03_Reverse_library_output/
seqs_rc.fna > 03_Combined_seqs.fna

$extract_seqs_by_sample_id.py –i 03_Combined_seqs.fna –o
03_seqs_by_sample.fasta –s NGS134.H23 –n

$cat 01B_seqs_by_sample.fasta 02_Combined_seqs.fna
03_seqs_by_sample.fasta > Combined_seqs.fna
```

Picking Operational Taxonomic Units (OTUs)

For the combined data sets, the workflow outlined for GS454-01A was run as separate scripts to enable the addition of a chimera checking step. After each step, the output(s) required for further scripts were moved into the working folder to ensure all required files were in the same working folder.

```
$pick_otus.py –i Combined_seqs.fna –o Picked_otus/

$pick_rep_set.py –i Combined_seqs_otus.txt –f Combined_seqs.fna –o
Rep_set.fna

$parallel_align_seqs_pynast.py –i Rep_set.fna –o Pynast_aligned/
```

A number of the more computationally intense scripts have alternative parallel scripts, which automatically run four processes at once and collate the results to look like those generated by the non-parallel variant of the script.

```
$identify_chimeric_seqs.py –m ChimeraSlayer –i rep_set_aligned.fasta –
a core_set_aligned.fasta –o Chimera_seqs.txt
```

QIIME's inbuilt ChimeraSlayer wrapper can be used to check all sequences for chimeras after alignment. The input file mist be in the same folder as the core_set_aligned.fasta reference file.

```
$filter_fasta.py –f rep_set_aligned.fasta –s Chimeric_seqs.txt –o Non_
chimeric_rep_set_aligned.fasta –n
```

The identified chimeric sequences are removed from the alignment prior to further analysis. The –n parameter ensures the sequences passed via –s are discarded, rather than keeping only those sequences.

```
$filter_alignment.py –i Non_chimeric_rep_set_aligned.fasta –m
lanemask_in_1s_and_0s –o filtered_alignment/

$parallel_assign_taxonomy_rdp.py –i Non_chimeric_rep_set_aligned_
pfiltered.fasta –o rdp_assigned_taxonomy/

$make_phylogeny.py –i Non_chimeric_rep_set_aligned_pfiltered.fasta –o
Combined_rep_set.tre

$make_otu_table.py –i Combined_seqs_otus.txt –e chimeric_seqs.txt –t
rdp_
assigned_taxonomy/Non_chimeric_rep_set_aligned_pfiltered_tax_assignmen
ts.
txt –o Combined_otu_table.biom
```

Including the chimeric sequences txt file with the –e parameter ensures these sequences are not included in the otu table.

```
$per_library_stats.py –i combined_otu_table.biom

$summarize_taxa_through_plots.py –i combined_otu_table.biom –o rdp_
assigned_taxonomy/ –m Combined_Metadata_forward.txt

$summarize_taxa_through_plots.py –i combined_otu_table.biom –o rdp_
assigned_taxonomy/ –m Combined_Metadata_forward.txt –c Species
```

Alpha diversity

```
$alpha_rarefaction.py –i Combined_otu_table.biom –m Combined_Metadata_
forward.txt –o alpha_rare/ -p alpha_params.txt –t Combined_rep_set.tre
–e 500 –a –O 4
```

Beta diversity

```
$beta_diversity_through_plots.py –i Combined_otu_table.biom –m
Combined_
Metadata_forward.txt –o beta_div/ -t Combined_rep_set.tre –e 500 –a –O
4
```

```
$jackknifed_beta_diversity.py –i Combined_otu_table.biom –t
Combined_rep_
set.tre –m Combined_Metadata_forward.txt –o jack_div/ -e 375 –a –O 4
```

Faecal bacteria beta diversity

```
$filter_taxa_from_otu_table.py –i Combined_otu_table.biom –o
Faecal_otu_table.biom –p
p__Bacteroidetes,p__Fibrobacteres,p__Firmicutes,
p__Fusobacteria,p__Proteobacteria –n c__Betaproteobacteria,
c__Deltaproteobacteria, c__Zetaproteobacteria
```

Filters an OTU table based on taxonomic metadata. The –p parameter allows a list of comma-separated taxa to be retained, while the –n parameter allows a list of taxa to be discarded. In this case, all *Bacteroidetes, Fibrobacteres, Firmicutes and Fusobacteria* were kept, while only taxonomic classes from *Proteobacteria* which are not *Betaproteobacteria*, *Deltaproteobacteria*, and *Zetaproteobacteria* were kept.

```
$filter_fasta.py –f Non_chimeric_rep_set_aligned_pfiltered.fasta –o
Faecal_rep_set.fasta –b Faecal_otu_table.biom
```

Filters sequences from the aligned and filtered file which are not contained in the new faecal bacteria otu table, to enable a new phylogenetic tree to be constructed with only the faecal bacteria.

```
$make_phylogeny.py –i Faecal_rep_set.fasta –o Faecal_rep_set.tre
```

```
$beta_diversity_through_plots.py –i Faecal_otu_table.biom –m Combined_
Metadata_forward.txt –o Faecal_beta_div/ -t Faecal_rep_set.tre –e 44 –
a
–O 4
```

```
$jackknifed_beta_diversity.py –i Faecal_otu_table.biom –m Combined_
Metadata_forward.txt –t Faecal_rep_set.tre–o Faecal_jack_div/ -e 30 –a
–O 4
```

Appendix II

QIIME Metadata files

GS45401_Metadata_forward

#Sample ID	BarcodeSequence	LinkerPrimerSequence	Species	ReversePrimer	Description
NGS001.B4.1	AGAG	AGAGTTTGATCCTGGCTCAG	Sheep	ACCGCGGCKGCTGGC	Sheep_1_composite
NGS002.B4.2	ACTC	AGAGTTTGATCCTGGCTCAG	Sheep	ACCGCGGCKGCTGGC	Sheep_2_composite
NGS003.B4.3	AGTG	AGAGTTTGATCCTGGCTCAG	Sheep	ACCGCGGCKGCTGGC	Sheep_3_composite_2step
NGS004.B4.4	ATAG	AGAGTTTGATCCTGGCTCAG	Sheep	ACCGCGGCKGCTGGC	Sheep_4_composite
NGS005.B4.5	ACAC	AGAGTTTGATCCTGGCTCAG	Cow	ACCGCGGCKGCTGGC	Cow_5_composite_2step
NGS006.B4.6	CACA	AGAGTTTGATCCTGGCTCAG	Cow	ACCGCGGCKGCTGGC	Cow_6_composite
NGS007.B4.7	CTCT	AGAGTTTGATCCTGGCTCAG	Cow	ACCGCGGCKGCTGGC	Cow_7_composite
NGS008.B4.8	CAGA	AGAGTTTGATCCTGGCTCAG	Cow	ACCGCGGCKGCTGGC	Cow_8_composite
NGS003.B4.9	CTGT	AGAGTTTGATCCTGGCTCAG	Sheep	ACCGCGGCKGCTGGC	Sheep_3_composite
NGS010.B4.10	ATGC	AGAGTTTGATCCTGGCTCAG	Duck	ACCGCGGCKGCTGGC	Duck_10_composite
NGS005.B4.11	GAGA	AGAGTTTGATCCTGGCTCAG	Cow	ACCGCGGCKGCTGGC	Cow_5_composite
NGS012.B4.12	GTGT	AGAGTTTGATCCTGGCTCAG	Swan	ACCGCGGCKGCTGGC	Swan_12_composite
NGS013.B4.13	GACA	AGAGTTTGATCCTGGCTCAG	Sewage	ACCGCGGCKGCTGGC	Human_13_sewage_C119
NGS014.B4.14	GTCT	AGAGTTTGATCCTGGCTCAG	Sewage	ACCGCGGCKGCTGGC	Human_14_sewage_CMB05123
NGS015.B4.15	GATC	AGAGTTTGATCCTGGCTCAG	Sewage	ACCGCGGCKGCTGGC	Human_15_sewage_CMB06668
NGS016.B4.16	TCTC	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_CMB120274
NGS017.B4.17	TGTG	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_CMB120322
NGS018.B4.18	TCTG	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_CMB120397
NGS019.B4.19	TCAC	AGAGTTTGATCCTGGCTCAG	Aquifer	ACCGCGGCKGCTGGC	Aquifer_1132
NGS020.B4.20	TGAG	AGAGTTTGATCCTGGCTCAG	Aquifer	ACCGCGGCKGCTGGC	Aquifer_10a_neat

GS45401_Metadata_reverse

#Sample ID	BarcodeSequence	LinkerPrimerSequence	Species	ReversePrimer	Description
NGS001.B4.1	AGAG	ACCGCGGCKGCTGGC	Sheep	AGAGTTTGATCCTGGCTCAG	Sheep_1_composite
NGS002.B4.2	ACTC	ACCGCGGCKGCTGGC	Sheep	AGAGTTTGATCCTGGCTCAG	Sheep_2_composite
NGS003.B4.3	AGTG	ACCGCGGCKGCTGGC	Sheep	AGAGTTTGATCCTGGCTCAG	Sheep_3_composite_2step
NGS004.B4.4	ATAG	ACCGCGGCKGCTGGC	Sheep	AGAGTTTGATCCTGGCTCAG	Sheep_4_composite
NGS005.B4.5	ACAC	ACCGCGGCKGCTGGC	Cow	AGAGTTTGATCCTGGCTCAG	Cow_5_composite_2step
NGS006.B4.6	CACA	ACCGCGGCKGCTGGC	Cow	AGAGTTTGATCCTGGCTCAG	Cow_6_composite
NGS007.B4.7	CTCT	ACCGCGGCKGCTGGC	Cow	AGAGTTTGATCCTGGCTCAG	Cow_7_composite
NGS008.B4.8	CAGA	ACCGCGGCKGCTGGC	Cow	AGAGTTTGATCCTGGCTCAG	Cow_8_composite
NGS003.B4.9	CTGT	ACCGCGGCKGCTGGC	Sheep	AGAGTTTGATCCTGGCTCAG	Sheep_3_composite
NGS010.B4.10	ATGC	ACCGCGGCKGCTGGC	Duck	AGAGTTTGATCCTGGCTCAG	Duck_10_composite
NGS005.B4.11	GAGA	ACCGCGGCKGCTGGC	Cow	AGAGTTTGATCCTGGCTCAG	Cow_5_composite
NGS012.B4.12	GTGT	ACCGCGGCKGCTGGC	Swan	AGAGTTTGATCCTGGCTCAG	Swan_12_composite
NGS013.B4.13	GACA	ACCGCGGCKGCTGGC	Sewage	AGAGTTTGATCCTGGCTCAG	Human_13_sewage_C119
NGS014.B4.14	GTCT	ACCGCGGCKGCTGGC	Sewage	AGAGTTTGATCCTGGCTCAG	Human_14_sewage_CMB05123
NGS015.B4.15	GATC	ACCGCGGCKGCTGGC	Sewage	AGAGTTTGATCCTGGCTCAG	Human_15_sewage_CMB06668
NGS016.B4.16	TCTC	ACCGCGGCKGCTGGC	Water	AGAGTTTGATCCTGGCTCAG	Water_CMB120274
NGS017.B4.17	TGTG	ACCGCGGCKGCTGGC	Water	AGAGTTTGATCCTGGCTCAG	Water_CMB120322
NGS018.B4.18	TCTG	ACCGCGGCKGCTGGC	Water	AGAGTTTGATCCTGGCTCAG	Water_CMB120397
NGS019.B4.19	TCAC	ACCGCGGCKGCTGGC	Aquifer	AGAGTTTGATCCTGGCTCAG	Aquifer_1132
NGS020.B4.20	TGAG	ACCGCGGCKGCTGGC	Aquifer	AGAGTTTGATCCTGGCTCAG	Aquifer_10a_neat

GS45402_Metadata_forward

#Sample ID	BarcodeSequence	LinkerPrimerSequence	Species	ReversePrimer	Description
NGS021.H2	TCACAGCA	AGAGTTTGATCCTGGCTCAG	Chicken	ACCGCGGCKGCTGGC	Chicken_21_composite
NGS028.H16	CGCTATGA	AGAGTTTGATCCTGGCTCAG	Chicken	ACCGCGGCKGCTGGC	Chicken_28_composite
NGS029.H17	GACACTAC	AGAGTTTGATCCTGGCTCAG	Chicken	ACCGCGGCKGCTGGC	Chiken_29_composite
NGS031.H3	GTAGCACT	AGAGTTTGATCCTGGCTCAG	Sheep	ACCGCGGCKGCTGGC	Sheep_31_composite
NGS038.H4	ATAGCGTC	AGAGTTTGATCCTGGCTCAG	Cow	ACCGCGGCKGCTGGC	Cow_38_composite
NGS057.H19	GACTGATC	AGAGTTTGATCCTGGCTCAG	Dog	ACCGCGGCKGCTGGC	Dog_57_composite
NGS059.H7	GTACGCAT	AGAGTTTGATCCTGGCTCAG	Dog	ACCGCGGCKGCTGGC	Dog_59_composite
NGS060.H21	CACTGTAG	AGAGTTTGATCCTGGCTCAG	Dog	ACCGCGGCKGCTGGC	Dog_60_composite
NGS061.H22	TGCATACG	AGAGTTTGATCCTGGCTCAG	Dog	ACCGCGGCKGCTGGC	Dog_61_composite
NGS081.H6	CTACGACA	AGAGTTTGATCCTGGCTCAG	Duck	ACCGCGGCKGCTGGC	Duck_81_composite
NGS083.H24	AGCATCAC	AGAGTTTGATCCTGGCTCAG	Duck	ACCGCGGCKGCTGGC	Duck_83_composite
NGS084.H25	TACGTGCA	AGAGTTTGATCCTGGCTCAG	Duck	ACCGCGGCKGCTGGC	Duck_84_composite
NGS102.H14	TGCTACAG	AGAGTTTGATCCTGGCTCAG	Chicken	ACCGCGGCKGCTGGC	Chicken_102_composite
NGS103.H15	AGCTAGTC	AGAGTTTGATCCTGGCTCAG	Chicken	ACCGCGGCKGCTGGC	Chicken_103_composite
NGS119.H12	TCAGTCGA	AGAGTTTGATCCTGGCTCAG	Human	ACCGCGGCKGCTGGC	Human_119_composite
NGS120.H13	ACAGTGCT	AGAGTTTGATCCTGGCTCAG	Human	ACCGCGGCKGCTGGC	Human_120_composite

GS45402_Metadata_reverse

#Sample ID	BarcodeSequence	LinkerPrimerSequence	Species	ReversePrimer	Description
NGS021.H2	TCACAGCA	ACCGCGGCKGCTGGC	Chicken	AGAGTTTGATCCTGGCTCAG	Chicken_21_composite
NGS028.H16	CGCTATGA	ACCGCGGCKGCTGGC	Chicken	AGAGTTTGATCCTGGCTCAG	Chicken_28_composite
NGS029.H17	GACACTAC	ACCGCGGCKGCTGGC	Chicken	AGAGTTTGATCCTGGCTCAG	Chiken_29_composite
NGS031.H3	GTAGCACT	ACCGCGGCKGCTGGC	Sheep	AGAGTTTGATCCTGGCTCAG	Sheep_31_composite
NGS038.H4	ATAGCGTC	ACCGCGGCKGCTGGC	Cow	AGAGTTTGATCCTGGCTCAG	Cow_38_composite
NGS057.H19	GACTGATC	ACCGCGGCKGCTGGC	Dog	AGAGTTTGATCCTGGCTCAG	Dog_57_composite
NGS059.H7	GTACGCAT	ACCGCGGCKGCTGGC	Dog	AGAGTTTGATCCTGGCTCAG	Dog_59_composite
NGS060.H21	CACTGTAG	ACCGCGGCKGCTGGC	Dog	AGAGTTTGATCCTGGCTCAG	Dog_60_composite
NGS061.H22	TGCATACG	ACCGCGGCKGCTGGC	Dog	AGAGTTTGATCCTGGCTCAG	Dog_61_composite
NGS081.H6	CTACGACA	ACCGCGGCKGCTGGC	Duck	AGAGTTTGATCCTGGCTCAG	Duck_81_composite
NGS083.H24	AGCATCAC	ACCGCGGCKGCTGGC	Duck	AGAGTTTGATCCTGGCTCAG	Duck_83_composite
NGS084.H25	TACGTGCA	ACCGCGGCKGCTGGC	Duck	AGAGTTTGATCCTGGCTCAG	Duck_84_composite
NGS102.H14	TGCTACAG	ACCGCGGCKGCTGGC	Chicken	AGAGTTTGATCCTGGCTCAG	Chicken_102_composite
NGS103.H15	AGCTAGTC	ACCGCGGCKGCTGGC	Chicken	AGAGTTTGATCCTGGCTCAG	Chicken_103_composite
NGS119.H12	TCAGTCGA	ACCGCGGCKGCTGGC	Human	AGAGTTTGATCCTGGCTCAG	Human_119_composite
NGS120.H13	ACAGTGCT	ACCGCGGCKGCTGGC	Human	AGAGTTTGATCCTGGCTCAG	Human_120_composite

GS45403_Metadata_forward

#Sample ID	BarcodeSequence	LinkerPrimerSequence	Species	ReversePrimer	Description
NGS046.H5	CTAGCTGA	AGAGTTTGATCCTGGCTCAG	Horse	ACCGCGGCKGCTGGC	Horse_46_composite
NGS048.H20	TACTGCGA	AGAGTTTGATCCTGGCTCAG	Pig	ACCGCGGCKGCTGGC	Pig_48_composite
NGS079.H8	ACATGCGT	AGAGTTTGATCCTGGCTCAG	Duck	ACCGCGGCKGCTGGC	Duck_79_composite
NGS114.H9	GCATGTAC	AGAGTTTGATCCTGGCTCAG	Pukeko	ACCGCGGCKGCTGGC	Pukeko_114_composite
NGS136.H10	ATACGTGC	AGAGTTTGATCCTGGCTCAG	Possum	ACCGCGGCKGCTGGC	Possum_136_composite
NGS137.H11	GCAGTATC	AGAGTTTGATCCTGGCTCAG	Alpaca	ACCGCGGCKGCTGGC	Alpaca_137_composite
NGS134.H23	CACGTATG	AGAGTTTGATCCTGGCTCAG	Aquifer	ACCGCGGCKGCTGGC	Aquifer_1152
NGS018.H7	GTACGCAT	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_18_CMB120397
NGS125.H26	CGCATGTA	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_125_CMB120346
NGS126.H27	GCGTACAT	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_126_CMB120351
NGS127.H28	ATGCACGT	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_127_CMB120354
NGS128.H29	TCGTAGTG	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_128_CMB120477
NGS129.H30	GTGCATAC	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_129_CMB120701
NGS130.H31	ACGTATGC	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_130_CMB120750
NGS131.H32	GTGACATC	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_131_CMB120751
NGS131.H12	TCAGTCGA	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_131_CMB120751

GS45403_Metadata_reverse

#Sample ID	BarcodeSequence	LinkerPrimerSequence	Species	ReversePrimer	Description
NGS046.H5	CTAGCTGA	ACCGCGGCKGCTGGC	Horse	AGAGTTTGATCCTGGCTCAG	Horse_46_composite
NGS048.H20	TACTGCGA	ACCGCGGCKGCTGGC	Pig	AGAGTTTGATCCTGGCTCAG	Pig_48_composite
NGS079.H8	ACATGCGT	ACCGCGGCKGCTGGC	Duck	AGAGTTTGATCCTGGCTCAG	Duck_79_composite
NGS114.H9	GCATGTAC	ACCGCGGCKGCTGGC	Pukeko	AGAGTTTGATCCTGGCTCAG	Pukeko_114_composite
NGS136.H10	ATACGTGC	ACCGCGGCKGCTGGC	Possum	AGAGTTTGATCCTGGCTCAG	Possum_136_composite
NGS137.H11	GCAGTATC	ACCGCGGCKGCTGGC	Alpaca	AGAGTTTGATCCTGGCTCAG	Alpaca_137_composite
NGS134.H23	CACGTATG	ACCGCGGCKGCTGGC	Aquifer	AGAGTTTGATCCTGGCTCAG	Aquifer_1152
NGS018.H7	GTACGCAT	ACCGCGGCKGCTGGC	Water	AGAGTTTGATCCTGGCTCAG	Water_18_CMB120397
NGS125.H26	CGCATGTA	ACCGCGGCKGCTGGC	Water	AGAGTTTGATCCTGGCTCAG	Water_125_CMB120346
NGS126.H27	GCGTACAT	ACCGCGGCKGCTGGC	Water	AGAGTTTGATCCTGGCTCAG	Water_126_CMB120351
NGS127.H28	ATGCACGT	ACCGCGGCKGCTGGC	Water	AGAGTTTGATCCTGGCTCAG	Water_127_CMB120354
NGS128.H29	TCGTAGTG	ACCGCGGCKGCTGGC	Water	AGAGTTTGATCCTGGCTCAG	Water_128_CMB120477
NGS129.H30	GTGCATAC	ACCGCGGCKGCTGGC	Water	AGAGTTTGATCCTGGCTCAG	Water_129_CMB120701
NGS130.H31	ACGTATGC	ACCGCGGCKGCTGGC	Water	AGAGTTTGATCCTGGCTCAG	Water_130_CMB120750
NGS131.H32	GTGACATC	ACCGCGGCKGCTGGC	Water	AGAGTTTGATCCTGGCTCAG	Water_131_CMB120751
NGS131.H12	TCAGTCGA	ACCGCGGCKGCTGGC	Water	AGAGTTTGATCCTGGCTCAG	Water_131_CMB120751

Combined_Metadata_forward

#Sample ID	BarcodeSequence	LinkerPrimerSequence	Species	ReversePrimer	Description
NGS001.B4.1	AGAG	AGAGTTTGATCCTGGCTCAG	Sheep	ACCGCGGCKGCTGGC	Sheep_1_composite
NGS002.B4.2	ACTC	AGAGTTTGATCCTGGCTCAG	Sheep	ACCGCGGCKGCTGGC	Sheep_2_composite
NGS003.B4.3	AGTG	AGAGTTTGATCCTGGCTCAG	Sheep	ACCGCGGCKGCTGGC	Sheep_3_composite_2step
NGS004.B4.4	ATAG	AGAGTTTGATCCTGGCTCAG	Sheep	ACCGCGGCKGCTGGC	Sheep_4_composite
NGS005.B4.5	ACAC	AGAGTTTGATCCTGGCTCAG	Cow	ACCGCGGCKGCTGGC	Cow_5_composite_2step
NGS006.B4.6	CACA	AGAGTTTGATCCTGGCTCAG	Cow	ACCGCGGCKGCTGGC	Cow_6_composite
NGS007.B4.7	CTCT	AGAGTTTGATCCTGGCTCAG	Cow	ACCGCGGCKGCTGGC	Cow_7_composite
NGS008.B4.8	CAGA	AGAGTTTGATCCTGGCTCAG	Cow	ACCGCGGCKGCTGGC	Cow_8_composite
NGS003.B4.9	CTGT	AGAGTTTGATCCTGGCTCAG	Sheep	ACCGCGGCKGCTGGC	Sheep_3_composite
NGS010.B4.10	ATGC	AGAGTTTGATCCTGGCTCAG	Duck	ACCGCGGCKGCTGGC	Duck_10_composite
NGS005.B4.11	GAGA	AGAGTTTGATCCTGGCTCAG	Cow	ACCGCGGCKGCTGGC	Cow_5_composite
NGS012.B4.12	GTGT	AGAGTTTGATCCTGGCTCAG	Swan	ACCGCGGCKGCTGGC	Swan_12_composite
NGS013.B4.13	GACA	AGAGTTTGATCCTGGCTCAG	Sewage	ACCGCGGCKGCTGGC	Human_13_sewage_C119
NGS014.B4.14	GTCT	AGAGTTTGATCCTGGCTCAG	Sewage	ACCGCGGCKGCTGGC	Human_14_sewage_CMB05123
NGS015.B4.15	GATC	AGAGTTTGATCCTGGCTCAG	Sewage	ACCGCGGCKGCTGGC	Human_15_sewage_CMB06668
NGS016.B4.16	TCTC	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_CMB120274
NGS017.B4.17	TGTG	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_CMB120322
NGS018.B4.18	TCTG	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_CMB120397
NGS019.B4.19	TCAC	AGAGTTTGATCCTGGCTCAG	Aquifer	ACCGCGGCKGCTGGC	Aquifer_1132
NGS020.B4.20	TGAG	AGAGTTTGATCCTGGCTCAG	Aquifer	ACCGCGGCKGCTGGC	Aquifer_10a_neat
NGS021.H2	TCACAGCA	AGAGTTTGATCCTGGCTCAG	Chicken	ACCGCGGCKGCTGGC	Chicken_21_composite
NGS028.H16	CGCTATGA	AGAGTTTGATCCTGGCTCAG	Chicken	ACCGCGGCKGCTGGC	Chicken_28_composite
NGS029.H17	GACACTAC	AGAGTTTGATCCTGGCTCAG	Chicken	ACCGCGGCKGCTGGC	Chiken_29_composite
NGS031.H3	GTAGCACT	AGAGTTTGATCCTGGCTCAG	Sheep	ACCGCGGCKGCTGGC	Sheep_31_composite
NGS038.H4	ATAGCGTC	AGAGTTTGATCCTGGCTCAG	Cow	ACCGCGGCKGCTGGC	Cow_38_composite
NGS057.H19	GACTGATC	AGAGTTTGATCCTGGCTCAG	Dog	ACCGCGGCKGCTGGC	Dog_57_composite
NGS059.H7	GTACGCAT	AGAGTTTGATCCTGGCTCAG	Dog	ACCGCGGCKGCTGGC	Dog_59_composite
NGS060.H21	CACTGTAG	AGAGTTTGATCCTGGCTCAG	Dog	ACCGCGGCKGCTGGC	Dog_60_composite
NGS061.H22	TGCATACG	AGAGTTTGATCCTGGCTCAG	Dog	ACCGCGGCKGCTGGC	Dog_61_composite
NGS081.H6	CTACGACA	AGAGTTTGATCCTGGCTCAG	Duck	ACCGCGGCKGCTGGC	Duck_81_composite
NGS083.H24	AGCATCAC	AGAGTTTGATCCTGGCTCAG	Duck	ACCGCGGCKGCTGGC	Duck_83_composite
NGS084.H25	TACGTGCA	AGAGTTTGATCCTGGCTCAG	Duck	ACCGCGGCKGCTGGC	Duck_84_composite
NGS102.H14	TGCTACAG	AGAGTTTGATCCTGGCTCAG	Chicken	ACCGCGGCKGCTGGC	Chicken_102_composite
NGS103.H15	AGCTAGTC	AGAGTTTGATCCTGGCTCAG	Chicken	ACCGCGGCKGCTGGC	Chicken_103_composite
NGS119.H12	TCAGTCGA	AGAGTTTGATCCTGGCTCAG	Human	ACCGCGGCKGCTGGC	Human_119_composite

Combined_Metadata_forward continued

NGS120.H13	ACAGTGCT	AGAGTTTGATCCTGGCTCAG	Human	ACCGCGGCKGCTGGC	Human_120_composite
NGS046.H5	CTAGCTGA	AGAGTTTGATCCTGGCTCAG	Horse	ACCGCGGCKGCTGGC	Horse_46_composite
NGS048.H20	TACTGCGA	AGAGTTTGATCCTGGCTCAG	Pig	ACCGCGGCKGCTGGC	Pig_48_composite
NGS079.H8	ACATGCGT	AGAGTTTGATCCTGGCTCAG	Duck	ACCGCGGCKGCTGGC	Duck_79_composite
NGS114.H9	GCATGTAC	AGAGTTTGATCCTGGCTCAG	Pukeko	ACCGCGGCKGCTGGC	Pukeko_114_composite
NGS136.H10	ATACGTGC	AGAGTTTGATCCTGGCTCAG	Possum	ACCGCGGCKGCTGGC	Possum_136_composite
NGS137.H11	GCAGTATC	AGAGTTTGATCCTGGCTCAG	Alpaca	ACCGCGGCKGCTGGC	Alpaca_137_composite
NGS134.H23	CACGTATG	AGAGTTTGATCCTGGCTCAG	Aquifer	ACCGCGGCKGCTGGC	Aquifer_1152
NGS018.H7	GTACGCAT	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_18_CMB120397
NGS125.H26	CGCATGTA	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_125_CMB120346
NGS126.H27	GCGTACAT	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_126_CMB120351
NGS127.H28	ATGCACGT	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_127_CMB120354
NGS128.H29	TCGTAGTG	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_128_CMB120477
NGS129.H30	GTGCATAC	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_129_CMB120701
NGS130.H31	ACGTATGC	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_130_CMB120750
NGS131.H32	GTGACATC	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_131_CMB120751
NGS131.H12	TCAGTCGA	AGAGTTTGATCCTGGCTCAG	Water	ACCGCGGCKGCTGGC	Water_131_CMB120751